Ghent University
Faculty of Bioscience Engineering
Department of Biotechnology

Expanding the portfolio of synthetic biology tools in *Saccharomyces cerevisiae* for the optimization of heterologous production pathways at the transcriptional and translational level

Thomas Decoene

Thesis submitted in fulfillment of the requirements
for the degree of Doctor (Ph.D.) in Applied Biological Sciences

Academic year 2017-2018

Examination committee

Prof. John Van Camp (Ghent University) (chairman)
Prof. Yves Briers (Ghent University) (secretary)
Prof. Els Van Damme (Ghent University)
Prof. Alain Goossens (Ghent University, VIB)
Prof. Ronnie Willaert (Free University of Brussels)

Supervisors

Prof. Marjan De Mey (Ghent University)
dr. Sofie De Maeseneire (Ghent University)

Dean

Prof. Marc Van Meirvenne

Rector

Prof. Rik Van de Walle

Ghent University
Faculty of Bioscience Engineering
Department of Biotechnology

Expanding the portfolio of synthetic biology tools in *Saccharomyces cerevisiae* for the optimization of heterologous production pathways at the transcriptional and translational level

Thomas Decoene

Thesis submitted in fulfillment of the requirements
for the degree of Doctor (Ph.D.) in Applied Biological Sciences

Academic year 2017-2018

Dutch translation of the title:

Uitbreiding van het portfolio aan synthetische biologie *tools* in *Saccharomyces cerevisiae* voor de optimalisatie van heterologe productie *pathways* op het transcriptionele en translationele niveau

To refer to this thesis:

Decoene, T., 2018. Expanding the portfolio of synthetic biology tools in *Saccharomyces cerevisiae* for the optimization of heterologous production pathways at the transcriptional and translational level. Ph.D. thesis, Ghent University.

Cover illustration:

vska (123RF)

ISBN number: 9789463570961

Thomas Decoene was supported by a fellowship of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

WOORD VOORAF

Dit is het dan, het schrijven van de laatste bladzijden van mijn *'boekske'*. Die vier jaar onderzoek neerpennen achtte ik vorig jaar nog onmogelijk en was hetgeen waar ik zwaar tegen opzag. Uiteindelijk ben ik wel heel tevreden dat mijn doctoraat hier eindelijk ligt. Werken aan een doctoraat was een zalige periode, waar ik de volledige vrijheid had om me bij te scholen op tal van vlakken. Echter was het ook vaak een lastige tijd, waar tegenslagen in experimenten en de zoveelste *'scoop'* ervoor zorgden dat ik me dikwijls afvroeg waar ik in godsnaam aan begonnen was. Dit doctoraat zou er dan ook nooit gelegen hebben zonder de aanwezigheid en hulp van een fantastische groep mensen, zowel op als buiten het labo, die ervoor gezorgd hebben dat ik de zin had om te blijven doorzetten. Bij deze wil ik hen hier dus heel graag bedanken!

Om te beginnen aan een doctoraat heb je natuurlijk iemand nodig die in je gelooft en je de komende vier jaar ten volle wil steunen, een mentor of een promoter zoals dat dan heet in officiële termen. Marjan, daarvoor wil ik jou in eerste plaats bedanken! Je hebt me de kans gegeven om met bakkersgist, toen ook voor jou nog een vrij onbekend organisme, te beginnen binnen de MEMO groep. Ik kreeg de volledige ruimte om nieuwe technieken binnen het veld van de synthetische biologie in gist uit te proberen en mijn kennis bij te schaven op nationale en internationale congressen. Voor dit vertrouwen ben ik je oprecht zeer dankbaar! Uiteraard ook een welgemeende dankjewel aan mijn co-promoter Sofie, de bakkersgist post-doc op het labo! Altijd kon ik bij je terecht met allerhande vragen en het grondig nalezen van mijn teksten apprecieer ik enorm! Daarnaast zorgde je vaak voor de amusante momenten in de MEMO bureau; zo was het altijd hilarisch wanneer het fameuze *'mannekes'* weer maar eens door de bureau galmde toen het boeltje er op stelten stond.

Dit brengt me dan ook naadloos bij het MEMO hoofdstuk. Als thesisstudent kwam ik terecht in een gedreven bende enthousiastelingen waar er naast wetenschappelijk onderzoek plaats was voor zeer veel ambiance. Deze goede sfeer heeft er immers grotendeels toe bijgedragen dat ik overtuigd was om hier te starten met een doctoraat. Mijn voorgangers waren het drietal Frederik, Gert en Pieter. Al tijdens het schrijven van ons IWT voorstel werden we ingewijd in de groep met de *'Russische avond'* georganiseerd door Frederik, een

legendarische avond die nog bij velen van ons in het geheugen gegrift staat. Ook Gert en Pieter waren er van in het begin bij en wil ik bedanken voor de input omtrent experimentele set-ups en data-analyse. Pieter bruiste van de zotste ideeën en wou als het enigszins mogelijk was zijn labowerk laten uitvoeren door robots. Gert lag mee aan de basis van het yUTR verhaal, [na een zoveelste *scoop*] '*Thomas, kga helpen uw doctoraat redden'*, of hoe bijgevolg een varia projectje is uitgedraaid op een mooi hoofdstuk en artikel, bedankt daarvoor! Op dat moment ook nog aanwezig waren Joeri, die als het ware volledige metabolische pathways uit zijn hoofd kende, en Gaspard, anti-Windows en me dus altijd maar vragen als ik geen Ubuntu cd'tje moest hebben. Samen met Bob en Brecht vormde ik dan de volgende lichting. Bob, veelal bezig met de laatste *'geek stuff'* (3D printers, zichzelf laten chippen, *etc.*) en zeer belezen omtrent de laatste synthetische biologie trends. Brecht, kortweg Paepe had altijd wel een ludiek verhaal in petto; '*De Groote Gaston*' en het scheidingsmisverstand van *'de Xavier'* zijn maar een paar voorbeelden uit het brede oeuvre. Als ik de voorbije vier jaar ook ooit één serieus woord met je gewisseld heb zal het veel geweest zijn. Uitspraken zoals '*Ja, mijn grijze plek is ook een biosensor'* waren van die hilarische momenten. Tevens denk ik nog vaak met een grote glimlach terug aan onze trip met Nico vorig jaar naar Singapore en Indonesië. Btw, laat me zeker weten wanneer Biosensor Centre Paepe uit de startblokken schiet! Intussen werd ook onze opvolging verzekerd. Tom, de stille harde werker. Samen hebben we toch enkele mooie systemen uitgewerkt in het labo om onze geliefde gastheer bakkersgist te temmen. Maarten VB, je mag dan wel de *'man van glas'* genoemd worden, toch wist je er in te slagen om een veel winnaar te zijn, getuige daarvan je overwinning op het schuttersparcours aan Dikkebusvijver, de topscore bij de bureaudarts en je recente NAR publicatie. David, altijd bereid om mijn West-Vlaamse zinsconstructies te vertalen en te ontleden, en me bijgevolg op linguïstisch vlak wat bij te scholen. Mol, het officieuze MEMO lid, labo feestpreases en *'drager der korte broeken'*. Verbazend hoe je er telkens in slaagde om alle laatste labogeroddel te weten te komen. Lien en Chiara, die de mannelijke hegemonie in de groep durfden te doorbreken. Bedankt ook voor de organisatie van het fantastische MEMO weekend in Parijs! De groep zou uiteraard niet hetzelfde zijn zonder Jo. Ook jij bent er al van bij het schrijven van onze projecten bij en zorgde ervoor dat we ons bleven focussen om die wetenschappelijke doelen te halen. Daarnaast was je altijd een amusante metgezel om wat subtiele stekjes te geven aan Sofie, wat dan meestal uitdraaide in een leuk verbaal post-doc gevechtje. Ondanks de zware jaren blijf ik ook jouw immense kracht en moed om te blijven

gaan bewonderen! *Last but not least* mogen we uiteraard het tweetal Wouter en Dries niet vergeten. Mister W, de grootmeester van het praktisch labowerk en de Koepuur jukebox. Al vanaf mijn thesis kon ik je lastig vallen met allerlei vragen omtrent moleculair en analytisch werk en trapte je Koepuur avonden op gang alsof het niets was. El Duchi, meermaals slaagde je erin om me het bloed van onder de nagels te halen door weer maar eens af te geven op de triestige gisten die toch niets konden of me uit te maken voor successupportertje. Samen met Wouter was je ook gepassioneerd door voetbal en sport, wat het altijd leuk maakte om onze visies te laten schijnen over dat prachtige doelpunt of die bijzondere koers.

Ook wil ik graag alle Glycodirect, BioPort en InBio collega's bedanken. Eén voor één fantastische mensen die allemaal bijdragen tot een formidabele sfeer op het labo wat het werken des te aangenamer maakt. Bewijs daarvan de vele post labo-activiteiten zoals de kerstfeestjes, het paasontbijt, de barbecues, labo-uitstappen, een pint pakken in de Koe en de verscheidene sportactiviteiten. Zeker op sportvlak heb ik me op het labo altijd ten volle kunnen uitleven! Dankjewel daarvoor aan Margo, Stevie, Jorick, Koen, Sylwia, Mol, Maarten VB en recentelijk Veerle, Sven, Jelle en Matthieu. Samen vormden we het Inbiyolo/CS Barcelona minivoetbalteam, een amusante ploeg waar het motto 'deelnemen is belangrijker dan winnen' echt van tel was. Vaste afsluiter was dan ook het doorspoelen van het verlies of het vieren van een zoet smakende (zeldzame) overwinning achteraf in het GUSB, altijd een leuk moment! Daarnaast vertegenwoordigden we het labo jaarlijks op de Mister T triatlon met een bende topatleten. Merci daarvoor aan Dries D, Tom V, Martijn, Griet, Magali, Margo, Sophie, Karel, Maarten D, Gert, Hannes en Stevie om er telkens een geslaagde namiddag van te maken. Ook werden er op zonnige lente – en zomeravonden met het labofietsteam menige kilometers afgemaald. Nooit zal ik onze legendarische fietstocht met Gert, Robin, Maarten D, Magali, Karel en Stevie naar Brugge vergeten! En naar het schijnt zouden we nog eens zoiets over doen richting Roeselare…

Een speciaal woordje van dank gaat ook uit naar de *Yeastpalace*. Vroeger nog met Lien en Isabelle, en vandaag de dag met het jongere geweld Tom, Mol, Veerle en Yatti, een zeer aangename plek waar er tussen het werken door veel gelachen werd. Zo denk ik maar terug aan Mol met zijn droog ijs bommetjes of het volle bak zetten van de radio wanneer er weer eens een goede schijf te horen was, zoals met Rammstein hé Isa. Hierbij uiteraard ook een bijzondere vermelding voor mijn twee thesisstudenten Nathalie en Yatti. Bedankt om elk een jaar lang met veel enthousiasme en inzet mee te draaien in mijn onderzoek! Verder wil

ik nog mijn dank uitspreken voor Gilles, altijd bereid om te helpen bij de laatste labo *issues* en er dagelijks in slagend om het labo organisatorisch draaiende te houden zodat alles voorradig was om de gewenste experimenten uit te voeren!

Tevens naast het labo wens ik nog een heleboel mensen die voor mij een belangrijke rol spelen te bedanken. Beginnen doe ik bij de zwemclub. Al vele jaren maak ik deel uit van een fantastische trainersgroep die zich iedere zaterdagvoormiddag inzet om de zwemmicrobe aan onze leden door te geven. Het lesgeven (en af en toe nog zelf getraind worden) is iedere keer een leuke afwisseling in de week waar ik veel voldoening uit haal! Niet te vergeten uiteraard zijn de goeie maten uit Roeselare. Erik en Mathijs, zelfs met iedere ochtend hetzelfde te mogen horen; '*Ah, ons doctoraatstudentje rolt ook nog een keer uit zijn bed*' toen ik nota bene al om kwart na acht op was, waren de drie jaar co-housing hier in Gent een machtige periode! Vaak denk ik nog terug aan de FIFA competities op PlayStation of de legendarische Champions League en WK matchen waar ons supportersgedrag soms zeer hevig oplaaide. Samen met de *'in-house'* vogelpiekcompetitie vergezeld door Evelyn en Lies en de tafelvoetbal leek ons appartement vaak op een leutig café. Merci ook aan Elias, Karel H en Mathijs! Onze jaarlijkse citytrips zijn altijd iets om naar uit te kijken op het einde van het jaar. Ook de menige fietskilometers en andere reizen die we al deden of de gewone pils avonden zijn telkens een amusante bedoening. Mathieu en Karel VH, steevast paraat om een burger te gaan eten en achteraf wat te zeveren bij een *'kupke'*. Tevens nooit te vergeten is onze Schotlandreis, waar we met een zalige bende op de West-Highland Way veel *leute* gemaakt hebben ondanks de vele regen en wind die getrotseerd moest worden.

Als laatste wil ik nog de familie bedanken en uiteraard het warme nest in Roeselare. Mama en papa, merci om mij alle kansen te geven, me met goede raad bij te staan en me te steunen in alle stappen die ik onderneem! Onze reis met het gezin naar Istanbul enkele jaren geleden was een toppertje waar ik nog graag aan terugdenk! Maarten en Hanne, en recentelijk uiteraard Sophie en Tijl, allemaal bijdragend tot de gezellige chaos die er heerst en de *heimat* een plaats maken waar het heel leuk vertoeven is wanneer ik er nog eens ben, bedankt!

Ziezo, na bijna tien jaar te hebben doorgebracht op het Boerekot sluit ik hier een formidabele periode af. Tijd voor het opzoeken van nieuwe horizonten. Tot in den draai!

Thomas

CONTENTS

ABBREVIATIONS 1

CHAPTER 1 INTRODUCTION AND OUTLINE 5

CHAPTER 2 STANDARDIZATION IN SYNTHETIC BIOLOGY: AN ENGINEERING DISCIPLINE COMING OF AGE 19

2.1 ABSTRACT 21

2.2 STANDARDIZATION AS DRIVING FORCE 22

2.3 STANDARDIZED SYNTHETIC BIOLOGY PARTS 24

2.4 STANDARDIZING ASSEMBLY 26

2.5 STANDARDIZING CHARACTERIZATION AND REPORTING 26

2.6 TOWARD MORE DATA SHARING AND FORWARD ENGINEERING 30

2.7 CONCLUSION 32

CHAPTER 3 MODULATING TRANSCRIPTION THROUGH DEVELOPMENT OF SEMI-SYNTHETIC YEAST CORE PROMOTERS 35

3.1 ABSTRACT 37

3.2 INTRODUCTION 38

3.3 MATERIAL AND METHODS 42

3.3.1 Strains and media 42

3.3.2 Plasmid construction 42

3.3.3 Fluorescence and absorbance measurements 44

3.3.4 Data analysis 45

3.4 RESULTS AND DISCUSSION 46

3.4.1 The minimal TEF1 core promoter 46

3.4.2 Random TEF1 core promoter library 48

3.4.3 yUGG, a method for one step, random assembly of an UAS library 52

3.5 CONCLUSION 56

CHAPTER 4 TOWARD PREDICTABLE 5'UTRs IN *SACCHAROMYCES CEREVISIAE*: DEVELOPMENT OF A yUTR CALCULATOR 59

4.1 ABSTRACT 61

4.2 INTRODUCTION 62

4.3 MATERIAL AND METHODS 65

4.3.1 Strains and media 65

4.3.2 Plasmid construction 65

4.3.3 In vivo fluorescence measurements 66

4.3.4 Model feature quantification 67
4.3.5 Partial least squares (PLS) regression 68
4.3.6 Search algorithm for de novo design of 5'UTRs 69
4.3.7 Cultivation of p-coumaric acid production strains 69
4.3.8 Detection and quantification of p-coumaric acid 70
4.3.9 Data analysis 70
4.4 RESULTS AND DISCUSSION 71
4.4.1 Development of the yUTR calculator 71
4.4.2 The yUTR calculator compared to the Dvir model 75
4.4.3 Universal applicability of the yUTR calculator 77
4.4.4 Protein coding sequence influence and reverse engineering 79
4.4.5 Proof of concept: reliable p-coumaric acid production 80
4.5 CONCLUSION 83
CHAPTER 5 CRITICAL EVALUATION OF MULTICISTRONIC GENE EXPRESSION IN *SACCHAROMYCES CEREVISIAE* 87
5.1 ABSTRACT 89
5.2 INTRODUCTION 90
5.3 MATERIAL AND METHODS 93
5.3.1 Strains and media 93
5.3.2 Construction of vectors for fluorescent protein transcription units 93
5.3.3 Characterization of synthetic T2A derivatives 94
5.3.4 Assessment of increasing consecutive T2A numbers 94
5.3.5 Fluorescence and absorbance measurements 95
5.3.6 Western Blotting 96
5.3.7 Data analysis 97
5.4 RESULTS AND DISCUSSION 98
5.4.1 Expanding the T2A palette 98
5.4.2 Genomic integration of T2A sequences at the URA3 locus 103
5.4.3 Tri – and quadcistronic gene expression at the URA3 locus 106
5.5 CONCLUSION 112
CHAPTER 6 METABOLIC ENGINEERING OF *SACCHAROMYCES CEREVISIAE* INTO A PLATFORM STRAIN FOR THE PRODUCTION OF FLAVONOIDS 115
6.1 ABSTRACT 117
6.2 INTRODUCTION 118
6.3 MATERIAL AND METHODS 124
6.3.1 Strains and media 124
6.3.2 Construction of expression vectors for flavonoid biosynthesis 124

6.3.3 Plasmid construction for gene knock-outs and CRISPR/Cas9 127
6.3.4 Strain construction 128
6.3.5 Cultivation of yeast production strains 131
6.3.6 Detection and quantification of flavonoids and intermediates 131
6.3.7 Data analysis 132
6.4 RESULTS AND DISCUSSION 133
6.4.1 Design of the flavonoid pathway 133
6.4.2 Evaluating p-coumaric acid production in S. cerevisiae 133
6.4.3 Effect of an enhanced malonyl-CoA pool on naringenin production 140
6.4.4 De novo production of naringenin: combining enhanced flows toward aromatic amino acids and malonyl-CoA 142
6.5 CONCLUSION 147
CHAPTER 7 GENERAL DISCUSSION AND OUTLOOK 149
APPENDICES 165
BIBLIOGRAPHY 239
SUMMARY 265
SAMENVATTING 271
CURRICULUM VITAE 279

ABBREVIATIONS

5'RACE	5' rapid amplification of cDNA ends
5'UTR	5' untranslated region
ADH1	alcohol dehydrogenase 1
BIA	benzylisoquinoline alkaloids
bp	base pair
CAD	computer-aided design
CAPS	N-cyclohexyl-3-aminopropanesulfonic acid
CDS	coding sequence
CDW	cell dry weight
CNN	convolutional neural network
CPEC	circular polymerase extension cloning
CPR	cytochrome P450 reductase
CRISPR	clustered regularly interspaced short palindromic repeats
CSM	complete supplement mixture
CUP1	metallothionein
CV	cross-validation
CYC1	cytochrome c isoform 1
DAHP	3-deoxy-D-arabino-7- heptulosonate 7-phosphate
DNA	deoxyribonucleic acid
E4P	erythrose-4-phosphate
EFE	ensemble free energy
EMOPEC	empirical model and oligos for protein expression changes
ENO1	enolase I
F2A	2A peptide of the foot-and-mouth disease virus
FACS	fluorescence activated cell sorting
FIT medium	Feed-In-Time medium
FP	fluorescence – fluorescent protein
GAL	galactokinase
GFP	green fluorescent protein
GG	Golden Gate
gRNA	guide ribonucleic acid
GSG	glycine-serine-glycine

HIS5	histidinol-phosphate aminotransferase
IC	initiation complex
iGEM (competition)	international genetically engineered machine (competition)
IRES	internal ribosome entry site
LASSO	least absolute shrinkage and selection operator
LB	lysogeny broth
LEU2	beta-isopropylmalate dehydrogenase
MCF	microbial cell factory
MIA	monoterpene indole alkaloids
MPA	measured protein abundance
mRNA	messenger ribonucleic acid
MTP	microtiter plate
NBT/BCIP	nitroblue tetrazolium/5-bromo-4-chloro-3-indolyl phosphate
NIST	national institute of standards and technology
OD	optical density
OD600	optical density measured at 600 nm
OLS	ordinary least square
Oof_uAUG	out-of-frame upstream start codon
ORF	open reading frame
ori	origin of replication
P2A	2A peptide of the *Porcine teschovirus-1*
PA	protein abundance
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PEP	phosphoenolpyruvate
Phe	phenylalanine
PIC	pre-initiation complex
PLS	partial least square
PODAC	protected oligonucleotide duplex assisted cloning
PoPS	polymerase per second
PPA	predicted protein abundance
PPP	pentose phosphate pathway
qPCR	quantitative polymerase chain reaction
R^2	coefficient of determination
RBS	ribosome binding site
RMSEP	root mean squared error of prediction
RNA	ribonucleic acid
RPL8A	ribosomal 60S subunit protein L8A

RSS	mRNA secondary structure
SAC6	fimbrin, actin-bundling protein
SBOL	synthetic biology open language
SD medium	synthetic defined medium
SDS-PAGE	sodium dodecyl sulfate – polyacrylamide gel electrophoresis
SNP	single nucleotide polymorphism
T2A	2A peptide of the *Thosea asigna* virus
TALEN	transcription activator-like effector nuclease
TDH3	glyceraldehyde-3-phosphate dehydrogenase
TEF1	translational elongation factor EF-1 alpha
TEF2	translational elongation factor EF-1 alpha
TFBS	transcription factor binding site
tGUO1	terminator from Guo *et al.* (1996) [1]
TIF6	translational initiation factor 6
TSS	transcription start site
TU	transcription unit
Tyr	tyrosine
UAS	upstream activating sequence
uAUG	upstream start codon
uORF	upstream open reading frame
UPLC	ultra performance liquid chromatography
URA3	orotidine-5'-phosphate decarboxylase
URS	upstream repressive sequence
VEGAS	versatile genetic assembly system
WT	wild-type
yECitrine	yeast enhanced citrine
yEGFP	yeast enhanced green fluorescent protein
yGG	yeast Golden Gate
YNB	yeast nitrogen base
YOGE	yeast oligo-mediated genome engineering
yUGG	yeast UAS Golden Gate
ZFN	zinc-finger nuclease

CHAPTER 1 INTRODUCTION AND OUTLINE

Today's transition to a bio-based economy is driven by a growing awareness of environmental problems, climate change and depletion of fossil fuels. This paradigm shift led to an increased attention for the development of industrially relevant, green production processes based on renewable resources. Industrial or white biotechnology, which uses micro-organisms and enzymes for the production of bulk chemicals, pharmaceutical compounds and food – and feed additives, plays herein a prominent role. The emerging potential of this field in the last decades was aided by metabolic engineering, creating microbial cell factories with economically feasible titers, yields and productivities. In recent years, this field further expanded toward systems metabolic engineering, by using tools and strategies of more novel research areas like systems biology and synthetic biology [2,3]. Eminent examples of successfully developed white biotechnology processes are the production of the antimalarial drug precursor artemisinic acid [4] (Amyris), the strong anti-oxidant resveratrol [5] (Evolva) and bulk chemicals such as 1,4-butanediol [6] (BioAmber and Genomatica). Recently, also the startup Antheia, Inc. was founded whose mission is to bring biotechnologically produced opioids like thebaine and hydrocodone on the market [7].

Transforming ordinary micro-organisms into robust cells for the industrial production of non-native metabolites is however still a challenging undertaking. It mostly requires the introduction of heterologous biosynthetic pathways into the host organism of choice. This often implicates a dramatic change in the tightly regulated host metabolism, causing unwanted side reactions, metabolic burden and growth deficiencies, altogether leading to a loss in productivity. In view of the plethora of techniques available for pathway assembly [8–13] and the ever increasing price drops in DNA synthesis, the biggest challenge in microbial cell engineering today is finding an optimal balance between the novel production pathway and the native metabolism.

Developing an appropriate cell factory typically starts with choosing the ideal microbial host for the industrial process. One of these interesting host organisms is the unicellular eukaryote *Saccharomyces cerevisiae*. *S. cerevisiae*, baker's yeast or brewer's yeast is already used for centuries by human mankind for the production of food and beverages. With the elucidation of its genome in 1996 [14], *S. cerevisiae* also became an important eukaryotic model organism for molecular biology research and for use in industrial processes. In this respect, *S. cerevisiae* has some inherent advantages, like post-translational modifications to support functional expression of plant enzymes, cell organelles for compartmentalization

of specific production pathways, and resistance against phages and low pH, which decreases the risk of contamination and increases tolerance to (fermentative) byproducts [15–17]. Furthermore, cell organelles like the endoplasmic reticulum and Golgi apparatus are ideal environments for the functional expression of membrane-bound P450 enzymes which typically appear in many pathways of secondary metabolites. Finally, baker's yeast is Generally Recognized As Safe (*i.e.* GRAS status) facilitating industrial process approval.

Though its well-known genetic background and optimal properties for usage in an industrial environment, the transformation of baker's yeast into a robust cell factory remains a time-, cost – and labor-intensive process which makes a further expansion of the synthetic biology toolbox vital. In this respect, efforts in the development of synthetic biology tools already fastened the pace of the yeast strain engineering process in the last decade. Tools for genome engineering purposes (*e.g.* Transcription activator-like effector nucleases (TALENs) [18] and CRISPR/Cas9 [19]), pathway construction (*e.g.* EasyClone [20] and the versatile genetic assembly system (VEGAS) [21]), gene expression regulation (*e.g.* promoter [22], 5'UTR [23] and terminator [24] libraries) and *in vivo* metabolite detection (*e.g.* RNA and protein-based biosensors [25,26]) are well-established examples of today's available techniques. An extensive overview of the currently available synthetic biology tools for the development of yeast cell factories is given in some great reviews by Jullesson *et al.*, Jensen *et al.* and Fletcher *et al.* [27–29]. Still, one of the ambitions of synthetic biology is to implement genetic modifications in biological systems by the usage of elementary engineering principles [27]. In this view, the potential of synthetic biology is not fully exploited at the moment as the field is lagging behind compared to other mature engineering disciplines, like electronics and the automotive industry, due to the lack of openness and well-documented standards. **Chapter 2** in this doctoral research gives as such a thorough insight in the current status of standardization, or the lack thereof, in the field of synthetic biology. With the ever increasing complexity of building biological systems and the associated expansion of synthetic biology tools, the need for standardization in this relatively young engineering discipline is high. All steps in the design-build-test workflow for strain development were evaluated on their standardization efforts. In addition, standardization principles were extended toward uniform data sharing. The whole led to a proposal for a complete data management life cycle for synthetic biology, enabling an efficient flow of information between researchers. The availability of the proposed standardized sets of data

and parts in combination with model-based and data-driven approaches is crucial for predictable and faster biological engineering.

Another important goal of synthetic biology is to use these standardized tools for the efficient harmonization of heterologous pathways and the native metabolism, as such enabling to speed up the strain engineering process. One essential factor herein is to develop methods that are able to predict the behavior of well-characterized regulatory parts and even whole genetic circuits in a microbial host. While great progress in the expansion of yeast synthetic biology tools has been made (see above), the yeast toolbox still lacks some techniques for the efficient regulation of gene transcription and translation, two essential control levels in living cells. Synthetic well-characterized parts for gene expression, predictive methods for the design of (heterologous) pathways and a better understanding of multicistronic gene expression are current gaps in the yeast engineering toolbox. Hence, a first objective of this Ph.D. dissertation was to focus on engineering methods for the regulation of transcription with short, non-native regulatory parts, the predictable effect of 5' untranslated regions (5'UTRs) on translation (*cfr.* RBS calculator in *E. coli* [30]) and the evaluation of multicistronic gene expression. All developed tools were evaluated by measuring fluorescent reporter proteins, an established validation approach in synthetic biology. A concise overview of the different topics discussed in this thesis, where a distinction is made between synthetic biology tools playing their role in either gene transcription or gene translation, is presented in Figure 1.1.

More specific in **Chapter 3**, promoter engineering strategies were used to expand the *S. cerevisiae* promoter toolbox. The core promoter sequence of the *TEF1* promoter was unraveled and a short functional core promoter was determined. This minimal regulatory sequence was used to create a core promoter library which can be used for altering transcription levels in *S. cerevisiae*. The library was evaluated for its influence on gene expression and was compared to commonly used yeast promoters. Furthermore, to expand the expression range of a given core promoter, a standardized one-step assembly method to incorporate single and multiple upstream activating sequences (UASs) was developed.

Improving yeast's translational regulation tools was achieved by developing forward engineering principles, leading to reliable strain development, and by evaluating the use of multicistronic pathways, allowing a reduction in the number of regulatory elements required. In this respect, **Chapter 4** handles about the predictable influence of 5'

untranslated regions (5'UTRs) on translation initiation. Based on an existing data set of yeast 5'UTR sequences, a partial least square (PLS) regression model that links 5'UTR features with protein abundance was constructed. Next, this sequence-function model was used for the design of 5'UTR sequences with user-defined translation efficiencies. The overall predictive capacity of this data driven method was evaluated in different transcriptional and translational contexts *in vivo*. This research line resulted in the 'yUTR calculator' that can design 5'UTR sequences with a diverse range of desired translation efficiencies. These results confirmed the great potential of data driven approaches for reliable pathway engineering in *S. cerevisiae*.

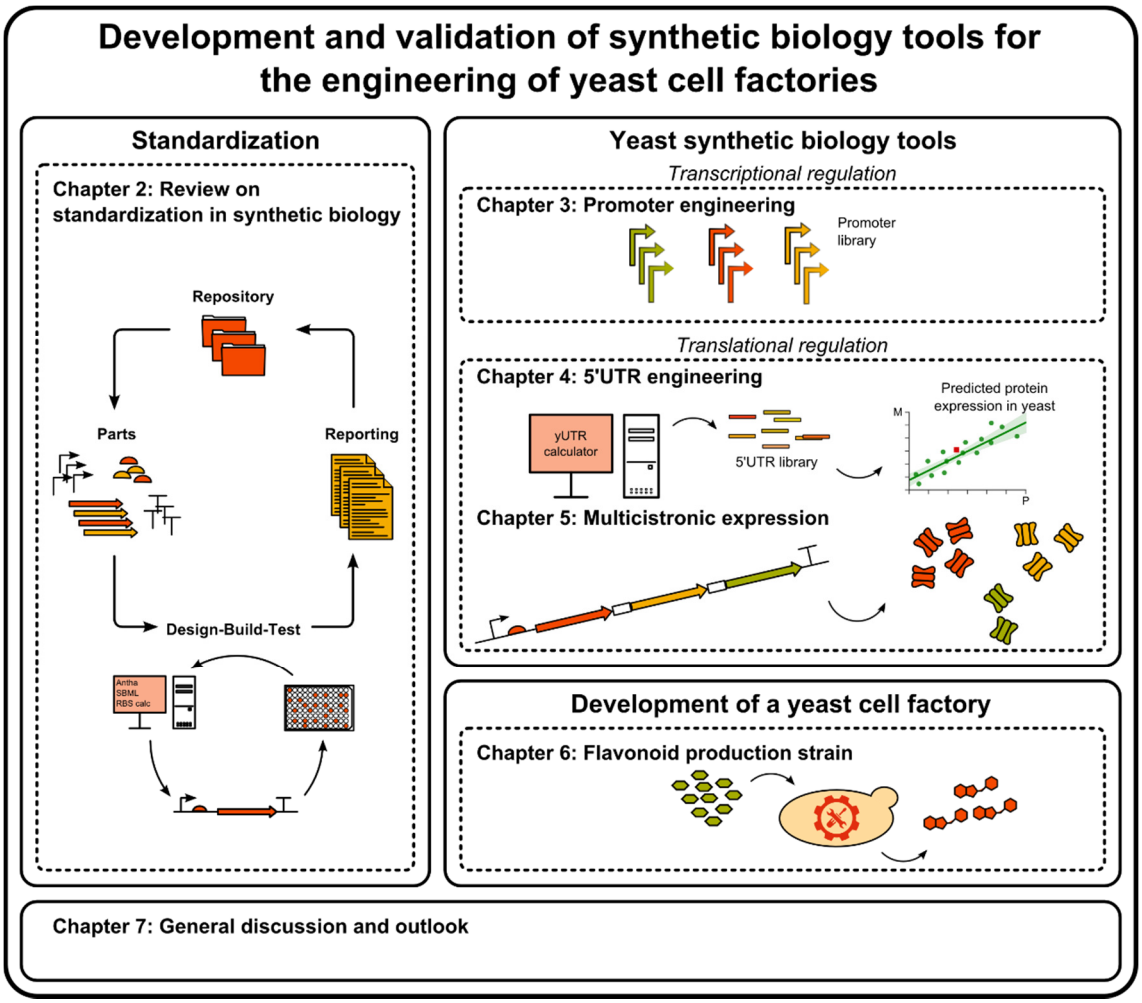

Figure 1.1: Overview of the different chapters discussed in this doctoral research project. The three main parts exist of: (1) an evaluation of standardization approaches in the synthetic biology field, (2) the development of tools to expand the yeast synthetic biology toolbox and (3) the construction of a yeast cell factory for flavonoid production.

The second tool for modulating gene expression at the translational level is described in **Chapter 5** where the ability of yeast to utilize multi – or polycistronic gene expression systems was assessed. Typically, every coding sequence (CDS) in eukaryotes is flanked at its 5' end by a promoter and at its 3' end by a terminator, which makes repeated use of these regulatory elements unavoidable in long pathways. This increases the risk of unwanted homologous recombination and thus strain instability, indicating the requirement of alternative expression units. Therefore, small 2A peptides causing ribosome skipping between two CDSs on a given mRNA were designed and characterized based on their splicing efficiency and protein expression capacity. Moreover, with the view on their application in large biosynthetic pathways, the effectiveness of 2A sequences in bi-, tri- and quadcistronic constructs in the genome was evaluated.

With view on the future applications of the developed synthetic biology tools in industrial biotechnology processes, a second goal of this Ph.D. dissertation was to transform *S. cerevisiae* into a robust host for the biosynthesis of phenylpropanoids (Figure 1.1). Phenylpropanoids, together with terpenoids and alkaloids, are generally known as secondary or specialty metabolites and are a class of compounds comprising over 200 000 different structures [31]. In general, these secondary metabolites are not linked to an organism's primary metabolism, which is essential for growth and reproduction, but merely play a role in defense and signaling mechanisms. As such, these molecules are mostly species dependent and have lots of biological activities which make them interesting target molecules for the pharmaceutical, cosmetic and food industry [31]. Typically, these secondary metabolites are naturally present in plants, fungi or other niches in the large, biodiverse Kingdom of Life. They are traditionally obtained by extraction from natural resources or via chemical synthesis. Since both methods have some inherent disadvantages like low yields, the use of hazardous solvents and harsh reaction conditions, the sustainable production of these specialty molecules in microbial cell factories is a worthy alternative to foresee them in sufficient amounts for human health applications. To this end, progress in the engineering of micro-organisms for the production of terpenoids, alkaloids and phenylpropanoids has been made in the last decades. Since many of the pathways of these compounds include P450-based enzymatic steps, which are more easily expressed in eukaryotic hosts, yeast is mostly the favorite organism to work with.

Terpenoids, also known as isoprenoids, form the largest group of plant secondary metabolites. They are formed out of the two universal C_5 precursor molecules isopentenyl pyrophosphate and dimethylallyl pyrophosphate, both synthesized via the mevalonate or the 2-C-methyl-D-erythritol-4-phosphate pathway [32]. Different metabolic engineering strategies in yeast and photosynthetic organisms already led to the biosynthesis of various terpenoids (extensively reviewed by Zhang *et al.* [33] and Arendt *et al.* [32]). Examples are artemisinic acid (25 g/l [34]), β-amyrin (36 mg/l [35]), patchoulol (42.1 mg/l [36]), miltiradiene (488mg/l [37]) and protopanaxadiol (1189 mg/l [38]).

Alkaloids are a class of nitrogen-containing metabolites typically derived from amino acids [31]. Benzylisoquinoline alkaloids (BIAs) and monoterpene indole alkaloids (MIAs), attain special attention because of their usefulness in the medical sector as analgesic, anticancer, antimicrobial and antiviral drugs [15]. Until recently, intermediate metabolites such as for example (R,S)-norlaudanosoline [39] were needed to produce BIAs or MIAs in a microbial host which made their production far from optimal. Especially the complex, long native biosynthetic pathways of alkaloids and the fact that not all genes in these pathways were unravelled made their *de novo* production challenging. Nevertheless, several breakthroughs the latest years in the discovery and engineering of novel enzymes, and the usage of enzyme-coupled biosensors for the optimization of production pathways made the *de novo* biosynthesis of alkaloids in yeast possible. For instance, the production of strictosidine (530 µg/l [40]), (S)-reticuline (80.6 µg/l [41], 19.2 µg/l [42]) and thebaine (6.4 µg/l [7]) was already demonstrated on lab-scale. Even though these titers are far too low for an economically viable production process, these studies are a starting point and show the great potential of future alkaloid fabrication by microbes.

Phenylpropanoids owe their name to the aromatic phenyl group and the propene tail obtained from cinnamic acid or p-coumaric acid [31]. Typically, phenylpropanoid biosynthesis is started from the amino acids phenylalanine or tyrosine. Amongst the group of phenylpropanoic compounds, flavonoids, consisting of over 6000 different structures [43], gain an increased attention in life science research due to their beneficial effects on human health. More specifically, it was shown that flavonoids have antibacterial, antiviral, anti-inflammatory, antioxidant and anticancer activities [44]. For several years, the production of these molecules in microbial platform organisms as *E. coli* and *S. cerevisiae* is on the rise. Flavanones like naringenin and pinocembrin, isoflavones like genistein and daidzein and

flavonols like kaempferol and quercetin were already successfully synthesized in *E. coli* or baker's yeast, mostly by supply of an intermediate (extensively reviewed by Trantas *et al.* and Pandey *et al.* [45,46]). Only very few studies describe the *de novo* production of flavonoids directly from glucose [47–54]. For example, to date only one study achieved to produce more than 100 mg/l naringenin under fed-batch conditions in a metabolically engineered yeast strain [48]. In this respect, transforming ordinary microbes in robust microbial cell factories able to produce flavonoids on an industrial scale is still a challenging undertaking, which formed the basis for further research in this dissertation. General strategies to overcome these limitations were recently reviewed by Delmulle *et al.* [55] who showed that engineering the host's native metabolism is a promising methodology to improve the phenylpropanoid precursor pools (*i.e.* phenylalanine, tyrosine and malonyl-CoA) and subsequent flavonoid production (Table 1.1). Yet, by our knowledge, no reports were published that exploited all three phenylpropanoid precursor pools to improve flavonoid production in yeast (Table 1.1). In particular, this means that improved phenylalanine and tyrosine pools, for respectively the plant and bacterial pathway toward p-coumaric acid, can be combined with an enhanced cytosolic malonyl-CoA pool. To this end, **Chapter 6** describes the strain engineering process to efficiently produce naringenin by optimizing fluxes toward tyrosine, phenylalanine and malonyl-CoA. First, competitive by-product formation and negative feedback inhibition was altered by gene knock-outs and protein engineering to enhance the tyrosine and phenylalanine pool, subsequently leading to better p-coumaric acid production. Next, malonyl-CoA supply was ensured by overexpression of an engineered acetyl-CoA carboxylase. This was evaluated by measuring naringenin titers when feeding with p-coumaric acid. As a final step, both approaches were combined for the production of naringenin directly from glucose. In addition, pioneering standardized assembly methods for large pathway construction in *S. cerevisiae* and CRISPR/Cas9 for genome editing purposes were used and validated. Furthermore, profitable flavonoid production also requires the fine-tuning of the introduced heterologous pathway which can be performed by *e.g.* selecting and/or engineering the right isoforms of the pathway enzymes, including multiple gene copies of rate-limiting enzymes and modulating enzyme expression by promoter, 5'UTR and terminator engineering [55]. Synthetic biology tools like these developed in the first part of this Ph.D. thesis could play herein a fundamental role. In this view, an initial test to reliably vary p-coumaric acid production by using the computational approach for predictive 5'UTR design developed in **Chapter 4** was performed.

In the final **Chapter 7**, the different techniques developed in this doctoral research to broaden the yeast synthetic biology toolbox and the metabolic engineering approaches followed for strain construction are evaluated. Moreover, perspectives to further apply these tools for the development of microbial cell factories, able to produce relevant compounds in an economically feasible manner, are given.

Table 1.1: Overview of strategies to improve the availability of precursor pools for the production of phenylpropanoids [5,48–50,56–78]. Table adapted from Delmulle *et al.* (2017) [55].

Precursor pool	Organism	Target genes [a]	Product titer	Fermentation type	Reference
Aromatic AAs	*E. coli*	Δ*tyrR, tyrA** ↑, *aroG** ↑	400 mg/l tyrosine	Batch	56
		Δ*tyrR, tyrA** ↑, *aroG** ↑, Δ*pheA*	893 mg/l tyrosine	Batch	57
		Δ*tyrR, tyrA** ↑, *aroG** ↑, *ppsA* ↑, *tktA* ↑	621 mg/l tyrosine	Batch	58
		*tyrA** ↑, *tyrB* ↑, *aroA* ↑, *aroB* ↑, *aroD* ↑, *aroE* ↑, *aroG** ↑, *aroL* ↑, *ppsA* ↑, *tktA* ↑	2169 mg/l tyrosine	Batch	59
		*pheA** ↑, *aroF* ↑	6720 mg/l phenylalanine [b]	Batch	60
		*tyrA** ↑, *aroG** ↑	100.64 mg/l 2S-naringenin [c]	Batch	61
		*pheA** ↑, *aroF* ↑	40.02 mg/l 2S-pinocembrin [c]	Batch	62
		*tyrA** ↑, *aroG** ↑,	41 mg/l genkwanin	Batch	63
	S. cerevisiae	*ARO4** ↑, *ARO7** ↑	235 mg/l resveratrol [c]	Batch	5
		ΔARO3, *ARO4** ↑, *ARO7** ↑	0.327 mmol tyrosine and phenylalanine $g^{-1}h^{-1}$	Chemostat	64
		ΔARO3, ARO4, ΔARO10, ΔPDC5, ΔPDC6*	54 mg/l naringenin	Batch	48
		*ARO4** ↑, *ARO7** ↑, *ΔARO10*, *aroL* ↑, *ΔPDC5*	1930 mg/l p-coumaric acid	Synthetic fed-batch	65
		*ARO4** ↑, *ARO7** ↑, *ΔARO10*, *aroL* ↑, *ΔPDC5*	1.6 mg/l naringenin	Synthetic fed-batch	49
		ARO4, PHA2*↓, *ΔARO10, ΔPDC5* [d]	84 mg/l naringenin	Batch	50
		*ARO4** ↑, *ARO7** ↑, *ΔARO10, TYR1* ↑, *ΔZWF1*	350 mg/l tyrosine	Batch	66
Malonyl-CoA	*E. coli*	*accAB* ↑, *fabF* ↑, *acs* ↑, Δ*ackA-pta*, Δ*adhE*	1280 mg/l phloroglucinol	Batch	67
		Δ*sdhA*, Δ*adhE*, Δ*brnQ*, Δ*citE*, ACC ↑, BPL ↑	215 mg/l naringenin	Batch	68
		ACC ↑, *MCR* ↑	1800 mg/l 3-hydroxypropionic acid	Batch	69
		acs ↑, *PlACC* ↑	429 mg/l pinocembrin	Batch	70
		Δ*fumC*, Δ*sucC*, *acc* ↑, *pgk* ↑, *pdh* ↑	474 mg/l naringenin	Batch with single feed	71

Precursor pool	Organism	Target genes [a]	Product titer	Fermentation type	Reference
Malonyl-CoA	*E. coli*	*fabF* ↑	59 mg/l pinosylvin	Batch	72
		fabF ↑	25.8 mg/l pinocembrin	Batch	73
		fabD ↓	91.31 mg/l naringenin	Batch	74
		adhE ↓, *fabF* ↓, *fabB* ↓, *fumC* ↓, *sucC* ↓	421.6 mg/l naringenin	Batch	75
		matB ↑, *matC* ↑	100.64 mg/l 2S-naringenin [e]	Batch	61
		matB ↑, *matC* ↑	40.02 mg/l 2S-pinocembrin [e]	Batch	62
	S. cerevisiae	*AAE13* ↑	3.5 mg/l resveratrol	Batch	76
		ACC1 ↑	554 mg/l 6-methylsalicylic acid	2 l bioreactor	77
		*ACC1** ↑	235 mg/l resveratrol [e]	Batch	5
		*ACC1** ↑	279 mg/l 3-hydroxypropionic acid	0.6 l bioreactor	78

[a] Δ: knock out, ↓: downregulation, *: mutation, ↑: overexpression

[b] This research used an L-tyrosine auxotrophic strain

[c] In combination with engineering of the malonyl-CoA precursor pool

[d] Galactose inducible naringenin pathway genes

[e] In combination with engineering of the aromatic amino acid precursor pool

CHAPTER 2 STANDARDIZATION IN SYNTHETIC BIOLOGY: AN ENGINEERING DISCIPLINE COMING OF AGE

2.1 ABSTRACT 21
2.2 STANDARDIZATION AS DRIVING FORCE 22
2.3 STANDARDIZED SYNTHETIC BIOLOGY PARTS 24
2.4 STANDARDIZING ASSEMBLY 26
2.5 STANDARDIZING CHARACTERIZATION AND REPORTING 26
2.6 TOWARD MORE DATA SHARING AND FORWARD ENGINEERING 30
2.7 CONCLUSION 32

Authors:

Thomas Decoene, Brecht De Paepe, Jo Maertens, Pieter Coussement, Gert Peters, Sofie De Maeseneire and Marjan De Mey

This chapter has been published as:

Decoene, T., De Paepe, B., Maertens, J., Coussement, P., Peters, G., De Maeseneire, S. L., and De Mey, M. (2017) Standardization in synthetic biology: an engineering discipline coming of age. *Crit. Rev. Biotechnol.* 1–10.

Author contributions:

All authors were involved in the conception and design of the paper. TD, JM, SDM and MDM drafted the manuscript. BDP designed the figures. All authors revised the manuscript critically.

2.1 ABSTRACT

Background: Leaping DNA read-and-write technologies, and extensive automation and miniaturization are radically transforming the field of biological experimentation by providing the tools that enable the cost-effective high-throughput required to address the enormous complexity of biological systems. However, standardization of the synthetic biology workflow has not kept abreast with dwindling technical and resource constraints, leading, for example, to the collection of multi-level and multi-omics big data sets that end up disconnected or remain under- or even unexploited.

Purpose: In this contribution, we critically evaluate the various efforts, and the (limited) success thereof, to introduce standards for defining, designing, assembling, characterizing and sharing synthetic biology parts. The causes for this success or the lack thereof, as well as possible solutions to overcome these, are discussed.

Conclusion: Akin to other engineering disciplines, extensive standardization will undoubtedly speed-up and reduce the cost of bioprocess development. In this respect, the further implementation of synthetic biology standards will be crucial for the field to redeem its promise, *i.e.*, to enable predictable forward engineering.

2.2 STANDARDIZATION AS DRIVING FORCE

Synthetic biology has come a long way since the introduction of the first chemical synthesis methods for DNA oligonucleotides and their assembly into larger DNA constructs. This rapidly advancing field has enabled the industrial biotechnological production of a wide range of bulk and fine chemicals from renewable resources not imaginable hitherto [5–7,34,79–81]. However, the development of the required microbial cell factories remains a long and labor-intensive undertaking with an uncertain outcome. In this respect, the lack of reliable, characterized and standardized biological parts for predictable strain engineering forms a major obstacle. Although well-defined parts have a strong track-record in more mature engineering disciplines, like in electronics, where automation and standardization with parts as resistors, capacitors and transistors massively contributed to success, the development of similar standardized parts for use in the field of synthetic biology is still in its infancy. Such parts and protocols, in combination with systems biology tools, will help to properly address the enormous cellular complexity [80,82,83] and contribute to the reproducibility in biological experimentation [84,85].

The awareness is indeed growing that extensive standardization is essential for the field to redeem its promise, *i.e.*, to enable predictable forward engineering. For example, elements for predictable heterologous gene expression are being constructed and extensively characterized [86–91], and tools like Cello [92], the Synthetic Biology Open Language (SBOL) [93,94], and Antha (URL: https://www.antha-lang.org/) were developed primarily to encourage design-oriented synthetic biology, lab automation and public reporting of part data. In addition, institutes like the National Institute of Standards and Technology (NIST) take lead to define and discuss standardization in synthetic biology, from DNA building block standards to documentation standards [95,96].

Despite these efforts, much remains to be done, *e.g.*, most 'standardized' synthetic biology parts are poorly and inconsistently characterized [86,87,97,98], and are influenced by the genetic and cellular context and the cellular environment, making them inappropriate for robust forward engineering in different cellular backgrounds or environmental conditions. This ineptness becomes ever more bothersome with the increasing complexity of genetic designs and the concurrent need for flexibility, standardized workflows and automation. In this contribution, we discuss the efforts already undertaken, or the lack thereof, and their

successes and shortcomings to introduce standards for defining, designing, assembling, characterizing and sharing synthetic biology parts (Figure 2.1).

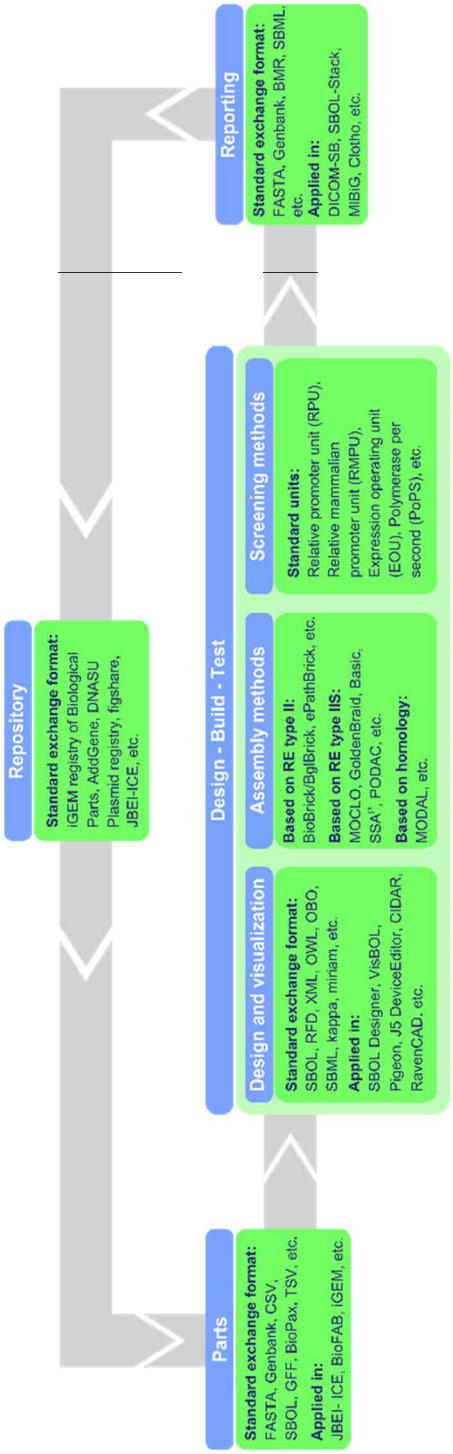

Figure 2.1: Non-exhaustive overview of current standardization efforts in synthetic biology.

2.3 STANDARDIZED SYNTHETIC BIOLOGY PARTS

Biological parts play a crucial role in the synthetic biology era. However, the lack of a clear way of defining such biological parts considerably complicates standardization. Previously, a standard biological part was defined by Canton *et al.* (2008) as "*a genetically encoded object that performs a biological function and that has been engineered to meet specified design or performance requirements*" [99]. Later, Lucks *et al.* (2008) additionally defined five key performance requirements of an optimal biological part, *i.e.,* independence, reliability, tunability, orthogonality and composability, which combined lead to the sixth equally important property, scalability [97]. Characterizing these six performance requirements would allow a more fundamental understanding and, as such, enable reliable forward engineering [97]. Yet, this dual definition, emphasizing the importance of both function and performance, still does not give synthetic biologists much to hold on to.

Thus far, parts seem to be typically defined based on functionality requirements. However, even in the well-known textbook organism *E. coli* the use of the term 'promoter' is in a way open for interpretation. Indeed, regions ranging from −57 till +4 have been denoted as sigma 70 promoters (Figure 2.2) [86,89,100–109]. In *S. cerevisiae* the term is even more untidily dealt with. Most commonly, transcriptional and translational control elements like the promoter *sensu stricto*, the 5' untranslated region (5'UTR) and the Kozak sequence are all captured in the term 'promoter' [23,110–112]. For example, in the recently described Yeast Toolkit from Lee *et al.* (2015), the 700 bps upstream of the start codon are featured as the promoter without any consideration of a 5'UTR or Kozak sequence [113]. In this context, the development of high-throughput methods for part identification, *e.g.*, via protein–DNA interactions such as *in vivo*-based ChIP-chip [114] and ChIP-seq methods [115] are expected to further boost the functional demarcation of biological parts.

Incorporating the performance criterions seems attractive, but does pose major problems [116]. For example, the classic abstraction of promoters as strictly transcriptional control elements and 5'UTRs as strictly translational control elements in *E. coli* is in a way ambiguous [86,89]. For instance, Salis *et al.* (2009) demonstrated that ribosome binding sites (RBS) with a predictive outcome can be designed, but that the performance of such an RBS is coding sequence dependent [30]. Though (some of) these effects were regarded evident for *E. coli*, in reality, it is hard to draw performance borders on the DNA level, *e.g.*, for promoters and RBSs. This recent change in opinion is substantiated by Kosuri *et al.* (2013), who

developed a screening method to determine the performance of novel biological part combinations, rather than relying on the predictive outcome of existing standardized parts [87]. Another example of such context effects is the impact of the chromosomal integration on expression, as demonstrated in *E. coli* [117,118], *B. subtilis* [119] and *S. cerevisiae* [120]. The development of transcriptional and translational insulator sequences to separate core elements from their genetic context [89,121–123] could be a (partial) solution to this problem, however, recent insights proved them not to be as generally functional as initially presumed [124]. To conclude, the independence criterion as proposed by Lucks *et al.* (2008) is utopian as other (neighbouring) elements will always affect the functionality of a DNA part [97]. Accordingly, synthetic biological parts which meet all criteria as defined by Lucks *et al.* (2008) may be nonexistent.

Despite abovementioned flaws, *i.e.*, the fuzzy demarcation of a DNA part and the influence of adjacent regions on a part's performance, and notwithstanding the immense complexity of a cell, DNA parts have been successfully applied for predictable pathway engineering. Recently, Nielsen *et al.* (2016) demonstrated that starting from well-characterized parts such reliable circuit design can be achieved, which, further finalized with extensive standardized screening and characterization methods to measure a circuit output, indicates real potential for future (forward) biological engineering [92].

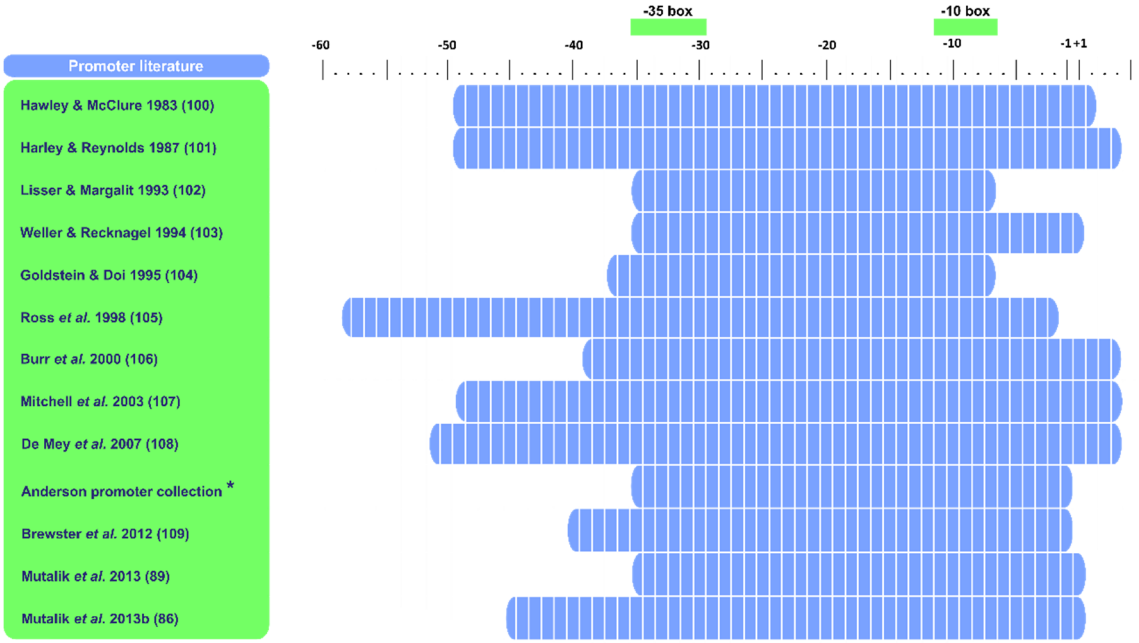

Figure 2.2: Examples of differently reported demarcations of a sigma70 promoter in *E. coli* [86,89,100–109] (*URL: http://parts.igem.org/Promoters/Catalog/Anderson).

2.4 STANDARDIZING ASSEMBLY

With the advent of synthetic biology, several novel DNA assembly methods and standards were reported [125,126]. These methods mainly use simple techniques redesigned to be applied on larger scale. To be able to scale, standardization is required in some form. Already 13 years ago, initial efforts resulted in the BioBricks and BglBricks design initiatives [127–129]. Since then several cloning standards have been published, *e.g.* various Golden Gate-based standards [9,13,21,130,131].

However, as some of these assembly methods require some sort of sequence modification, for example sticky ends necessary for base pairing in Golden Gate assembly or the lack of secondary structures at the end of DNA parts for Gibson assembly, parts can often only be used within a certain assembly standard, which hampers their interchangeability. In this respect, reoccurring assembly scars are particularly undesired [132], *e.g.*, as hotspots for recombination. In response, several scarless and sequence-independent methods have been developed in recent years, such as DATEL [133], Twin-Primer Assembly (TPA) [134] and CasHRA, *i.e.,* a Cas9 dependent assembly technique [135].

Still, in view of the huge price drops in DNA synthesis, demonstrated by the appearance of 'DNA factories' and full automatic DNA assembly lines, assembly concerns are no longer the most prominent issues.

2.5 STANDARDIZING CHARACTERIZATION AND REPORTING

Thorough part characterization is often performed within research groups with the goal to rationally use and reuse designed parts. However, as the goals of research groups are different, so are the protocols that are being used for part characterization. Different circumstances and measured parameters render this data field very fragmented. However, standardized part characterization and reporting is not only essential to reduce performance variability, but also to facilitate and promote interlab reusability of the data and ultimately of the parts themselves. To this end, for example, standardized protocols for fluorescent-activated cell sorting (FACS) and GFP quantification (URL: http://openwetware.org/wiki/Main_Page) have been introduced. Moreover, to assure the quality of the data collected, ring tests are increasingly being used. For example, in last year's iGEM competition an Interlab Measurement Kit was distributed to quantify and compare fluorescence measurements across different teams [136].

The evolution toward more standardization in the process of part characterization is also supported by the successful penetration of more high-throughput lab automation equipment such as liquid handling robotics and microfluidic systems [137–139]. Lab automation offers an opportunity for the development of standardized workflows from concept to lab, steering all steps in between designing, screening, analyzing and interpreting the experimental outcome [137,140,141]. Moreover, this automation puts an end to the main source of variability in parts characterization, *i.e.*, 'human practice' [136]. This evolution is also supported by the development of tools like Antha (URL: https://www.antha-lang.org), a high-level programming language for biology developed to build reproducible, scalable workflows drawn on reusable elements; SBOL [93,94], the synthetic biology standard which also supports development of genetic design automation software [142], and recently Cello, specifically developed for reliable, automated circuit design [92].

To date, to characterize part performance *in vivo*, fluorescent proteins are still the standard tool [143], despite inherent disadvantages like maturation times, stability, leakage to the medium, different available isoforms and oxygen dependence [144,145]. However, these fluorescent proteins often do not allow to quantify the performance of a synthetic biology part itself, rather of multiple individual parts. The latter is particularly problematic in view of the construction of ever more complicated multi-part biological devices and genetic circuitry. In response, techniques such as qPCR, RNA IMAGEtags reporters [146] and RNA aptamers [147,148] are being applied to discern the contribution of various parts on, *e.g.*, gene expression. Such multi-level and multi-omics characterization of part performance is on the up and complies well with advances in the field of systems biology.

Another approach to eliminate the possible interference of the multi-level regulation *in vivo*, is *in vitro* characterization [149]. In this way, synthetic biology parts can be characterized in well-controlled reaction conditions. Moreover, these cell-free systems can be integrated in high-throughput systems and avoid classic *in vivo* molecular techniques such as transformations and plasmid recovery, which speeds up the characterization process considerably [150]. In this way, complex genetic circuitry to be used *in vivo* can be more successfully developed, by debugging it, beforehand, *in vitro*. For example, several cell-free systems expressing synthetic circuits were built from *in vitro* characterized parts [151–153].

The wide variety of distinct measures, absolute and relative, each with their own specific characteristics, that can be found in literature is another major problem on the road to standardizing part characterization [92,98,99]. For example, to evaluate the transcription initiation rate of various promoters the measures 'specific fluorescence' [154], 'specific productivity' [155], 'relative promoter activity' [98,156], 'fluorescence' [111,157], 'promoter strength' [22,158], *etc.* are being used (Figure 2.3). Efforts to standardize these measures for reporting the performance of a synthetic biology part are illustrated by the introduction of standard units such as polymerase per second (PoPS). However, their applicability is hampered by the troublesome experimental setup, especially when characterized in different environmental conditions [98].

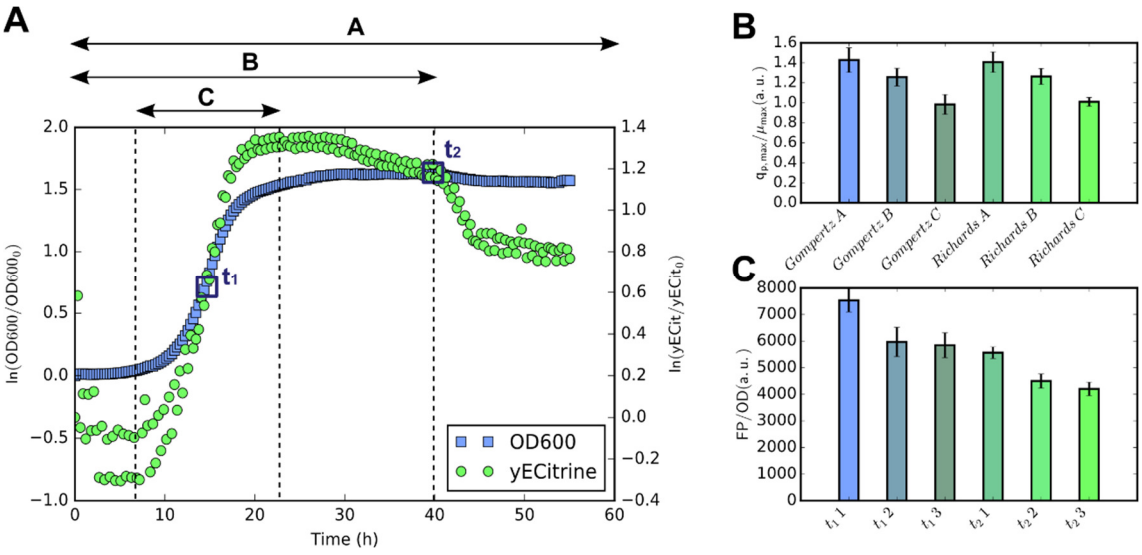

Figure 2.3: Overview of different measures used to characterize DNA parts. (A) Data from a growth experiment with *S. cerevisiae* SY992 (Euroscarf culture collection) measured every 15 minutes in a Tecan Infinite M200 plate reader. Square blue dots represent the optical density (OD) measured at 600 nm and green circular dots represent the fluorescence (FP) of a yECitrine reporter (pKT140, Euroscarf culture collection) controlled by the native CYC1 promoter (-287 to -1) [159], measured at 500 nm (excitation) and 540 nm (emission). (B) Impact of the time window on the fluorescence measure $q_{p,max}/\mu_{,max}$ determined by continuous growth fit models like the Gompertz and Richards fit [160]. Time window A; the complete growth and fluorescence curve is taken into account, inclusive the death phase. Time window B; the model is fit against the growth and fluorescence curve without accounting the death phase. Time window C; the model is fit against the exponential part of the growth and fluorescence curve. (C) Impact of timing (t1 and t2), and background correction on the endpoint measure of specific fluorescence FP/OD. Situation 1; FP/OD without any background correction for the wild type and the medium. Situation 2; $(FP-FP_{WT})/(OD-OD_{med})$ where a background correction for the wild type (WT) and the medium (med) is taken into account. Situation 3; $(FP/OD)-(FP/OD)_{WT}$ where the specific fluorescence of the wild type is used as background correction.

Synthetic biology parts that perform predictably and robustly under a wide array of environmental and genetic conditions are crucial for the success of synthetic and industrial biotechnology, *e.g.*, during scaling-up from perfectly controlled environments encountered at laboratory scale to the often harsh and fickle conditions encountered at industrial scale. In this respect, the numerous delays in the development of industrial biotechnology processes resulted in a loss of confidence between investors and industrial biotechnology in recent years. For instance, the launch of Evolva and Cargill's stevia is delayed until 2018 due to an 'unsatisfactorily performing yeast production host with lower than expected production yields and consequently higher than expected COGS' [161]. In general, the highly non-linear behavior and evolutionary nature of micro-organisms adversely affect robustness. For example, specific synthetic biology parts (and specific combinations thereof) may inflict severe cellular stress, dysfunction and even cell death [162], cells may evolutionary adapt themselves to the observed environmental conditions [163], *etc*. To enhance robustness, orthogonal expression systems [162] and genetic circuitry [164] can be implemented. In addition, model-based approaches such as bifurcation analysis have proven to be useful to evaluate, in advance, the robustness of heterologous pathway designs [165].

Part characterization constitutes thus an inherent trade-off between proven robustness and predictability of the -to be developed- engineered biological part, device or system and the analytical efforts required to demonstrate this under the relevant genetic and environmental contexts beforehand. Even though, it is often hard to anticipate which contexts will be relevant during the development of an industrial biological process.

Subsequently, the data collected on synthetic biology parts has to be reported preferably in a publically accepted and standardized format to maximize usability. Currently however, very diverse ways of data representation are being used, impeding the reuse of the evaluated parts. In response, to ensure an effective flow of information between researchers, the implementation of minimum information requirements and exchange format standards is vital. Minimum information requirements provide standardized specifications on what information about the experiment (metadata) is deemed critical to be reported in order to enhance the usability and reusability of the collected research data. Such minimal information requirements have been defined for genome and metagenome sequences (MIGS and MIMS) [166], microarray and proteomic experiments (MIQE [167], MIAMET

[168], MIAME [169] and MIAPE [170]), *etc*. However, such minimal information standards are far from being generally applied, as demonstrated by Chavez *et al.* (2016), who pinpointed that critical experimental settings of plate reader assays either vary between laboratories or are not reported, suggesting widespread reproducibility issues [171]. Standard exchange formats, *e.g.*, based on semantic web ontologies, on the other ensure easy data sharing and simultaneously assures data usability for computer-aided design tools. In this regard, the recent development of SyBiOnt [172], which is an application ontology that facilitates the modeling of information about biological parts and their relationships, is an important breakthrough, together with DICOM-SB [173], a new representation and communication standard specifically intended for biological part characterization.

2.6 TOWARD MORE DATA SHARING AND FORWARD ENGINEERING

The final step to ensure that the massive amounts of data generated on synthetic biology parts and devices do not end up disconnected or remain under – or even unexploited is standardization in data sharing (Figure 2.4). To this end, various data registries and repositories of parts and devices have already been established (Figure 2.1) and are cured regularly. Noteworthy examples are the Virtual Parts Repository (URL: http://sbol.ncl.ac.uk:8081/), the Registry of Standard Biological Parts (URL: http://parts.igem.org/), the Joint BioEnergy Institute's Inventory of Composable Elements (JBEI-ICE; URL: https://acs-registry.jbei.org), the Standard European Vector Architecture 2.0 database (SEVA-DB 2.0, URL: http://seva.cnb.csic.es/ [126]) and newly the Plant Associated and Environmental Microbes Database (PAMDB; URL: http://genome.ppws.vt.edu/cgi-bin/MLST/home.pl [174]). Some repositories, such as the Registry of Standard Biological Parts have also implemented quality control checks.

Data sharing is essential to accelerate progress, to reduce redundant efforts, to improve reproducibility and to allow reuse of past work. However, efforts to create openness by enlarging access to characterized parts for synthetic biology are fragmentary and hardly successful. In this respect, more will be needed for scientists to systematically deposit complete information on their parts and data than merely journal encouragements, *e.g.*, by ACS Synthetic Biology [175] and Molecular Biology Of the Cell (MBOC) (URL: http://www.ascb.org/). Despite such encouragements, the sequence information provided

in some synthetic biology publications is inadequate. Although most publications provide exhaustive descriptions of the methods used, obtaining full sequence information remains a daunting and sometimes impossible task [176]. To date, access to part (performance) data is often purposely withhold both by academia and by industry with a view to the development of a sustainable competitive advantage, whether or not by the development of intellectual property. Akin to the open access policy of various funding agencies, *e.g.*, green open access as required by the National Science Foundation, initiatives should be taken to encourage data and part sharing. For example, in the EU Framework Program Horizon 2020 for Research and Innovation, participants are required to develop a data management plan to ensure that the data generated are maximally exploited. In addition, community driven initiatives like FigShare (URL: https://figshare.com/) and Open Science Framework (URL: https://osf.io/) in general are getting more and more traction despite the fact that they are often subject to non-standardized reporting.

Besides, the development of a GenBank counterpart for biological parts and devices, *i.e.*, a unique and persistent identifier, will be crucial to deliver data consistency and easy access. Such identifiers should additionally allow to build a part's pedigree. Preliminary efforts are the establishment of the Inventory of Composable Elements (ICE), an open source registry software and platform for managing information about biological parts by the Joint BioEnergy Institute.

The availability of such multi-omics big data sets in combination with model-based and data-driven methods, such as machine learning techniques, are vital to properly handle the enormous complexity of biological systems and will undoubtedly contribute to the success of synthetic biology in the forward engineering of biological systems. For an increasing number of synthetic biology parts, *e.g.,* for promoters [108,177,178], ribosome binding sites [30,179–181], terminators [91] and riboswitches [182], model-based approaches have been developed in recent years to predict part performance, and in some cases, even to enhance our understanding of their mode of operation. Furthermore, the development of complex genetic circuitry, composed of multiple pre-characterized synthetic biology parts, with a desired dynamic behavior strongly relies on mathematical modelling and simulation [92,174,183–186]. In this respect, the use and further improvement of models will lead to more accurate predictions, which will reduce the number of design candidates and will lead to the

faster development of new synthetic circuits with a desired behavior, as such speeding-up and reducing the cost of bioprocess development.

2.7 CONCLUSION

Progress in industry and in academic research has often been impeded by the lack of reproducibility in biological experimentation due to the inadequate enforcement of standardized data processing workflows and to the lack of information required to reproduce the experiment.

To date, these threats are partially withholding synthetic biology from revolutionizing biological engineering. In this respect, the field is confronted by the lack of openness, an excrescence of so called standards and/or the limited success thereof for the various aspects of the synthetic biology workflow, *i.e.*, design, assembly, characterization, and data sharing, all jeopardizing the success of forward engineering in biological experimentation. Nonetheless, supported by the growing awareness of the usefulness of standards and in line with extensive miniaturization and automation, the evolution toward more extensive standardization is inevitable. However, to promote openness and streamline inter-laboratory standardization much remains to be done. To this end, also further compelling initiatives by journals and funding agencies will likely be necessary.

Crucial for the predictability in biological engineering, standardization in synthetic biology will ultimately contribute to the success of numerous disciplines, not in the least industrial biotechnology.

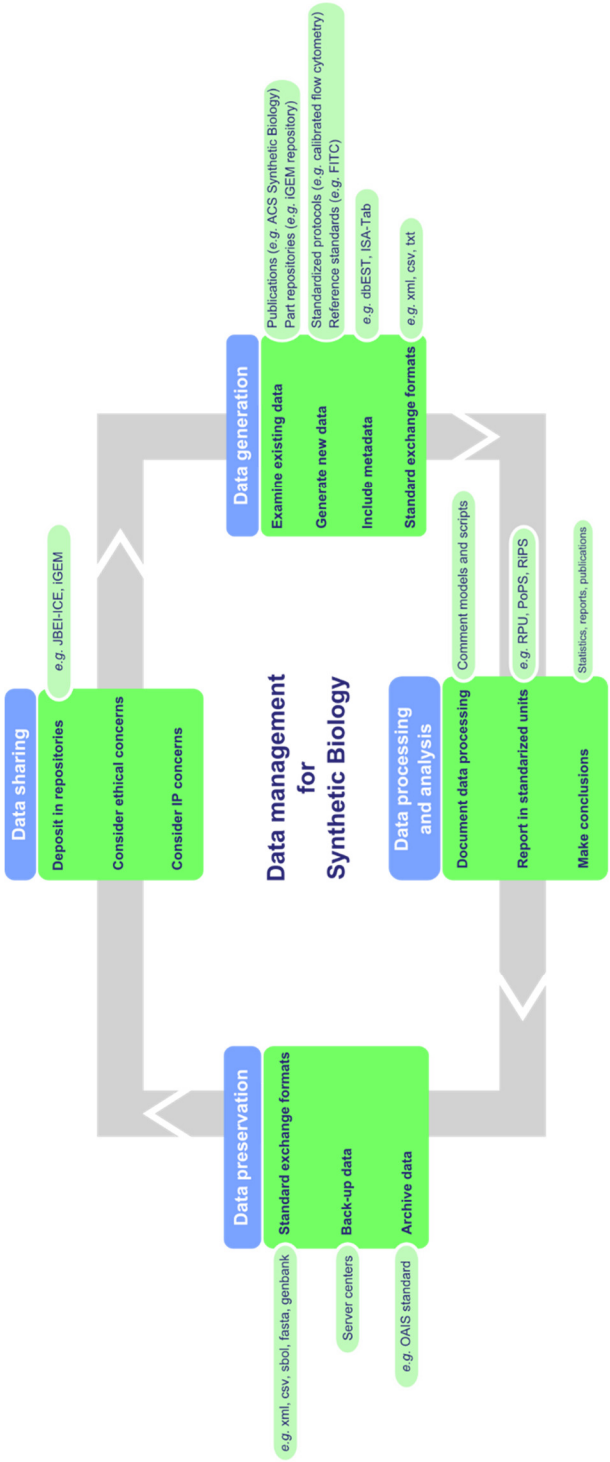

Figure 2.4: Data management life cycle for synthetic biology. To assure an efficient flow of information between researchers, an adequate data management plan is unavoidable. Important is to cover uniform exchange formats and metadata so it is possible to unambiguously interpret biological part data. In addition, encouraging more data sharing of parts is vital to provide input for in silico genetic circuit design by modeling. Abbreviations: FITC, fluorescein isothiocyanate; RPU, Relative Promoter Unit; PoPS, Polymerases per second; RiPS, Ribosomes per second; OAIS, Open Archival Information System; IP, Intellectual Property.

CHAPTER 3 MODULATING TRANSCRIPTION THROUGH DEVELOPMENT OF SEMI-SYNTHETIC YEAST CORE PROMOTERS

3.1 ABSTRACT 37
3.2 INTRODUCTION 38
3.3 MATERIAL AND METHODS 42
3.3.1 Strains and media 42
3.3.2 Plasmid construction 42
3.3.3 Fluorescence and absorbance measurements 44
3.3.4 Data analysis 45
3.4 RESULTS AND DISCUSSION 46
3.4.1 The minimal TEF1 core promoter 46
3.4.2 Random TEF1 core promoter library 48
3.4.3 yUGG, a method for one step, random assembly of an UAS library 52
3.5 CONCLUSION 56

Authors:

Thomas Decoene, Nathalie Cuypers, Sofie De Maeseneire and Marjan De Mey

Publication status:

Unpublished

Author contributions:

TD, SDM and MDM were involved in the conception, design and writing of the manuscript. Experiments were performed by TD and NC. TD performed the data analysis and interpretation of the results.

3.1 ABSTRACT

Altering gene expression regulation by promoter engineering is a very effective way to fine-tune heterologous pathways in eukaryotic hosts. Typically, pathway building approaches in yeast still use a limited set of long, native promoters. With the today's introduction of longer and more complex pathways, an expansion of this synthetic biology toolbox is necessary. In this study we elucidated the core promoter structure of the well-characterized yeast *TEF1* promoter and determined the minimal length needed for sufficient protein expression. Furthermore, this minimal core promoter sequence was used for the creation of a promoter library covering different expression strengths. This resulted in a group of short, 69 bp promoters with a 4.0-fold expression range and one exemplar that was doubled in activity compared to the native *ADH1* and *CYC1* promoters. Additionally, as it was earlier described that the protein expression range could be broadened by upstream activating sequences (UASs), we developed a standardized method called yUGG for the random integration of single and multiple UASs in front of short yeast promoters. With this approach, multiple and different UAS elements were added in a single one-pot assembly step to a truncated *TEF1* promoter, which contributed to further expression variation. As such, these results indicate the potential of standardized methods in yeast promoter engineering and the suitability of short yeast core promoters, either in an individual context or combined with UAS elements, for metabolic engineering applications.

3.2 INTRODUCTION

The yeast *Saccharomyces cerevisiae* serves as an ideal platform organism for the economically viable production of bulk and fine chemicals [27,187]. This however requires the introduction of heterologous metabolic pathways and the fine-tuning of gene expression to find an optimal balance within the production pathway, and between the host's native metabolism and the imbedded pathway. One effective way to alter and optimize metabolic pathways in yeast is gene expression regulation at the level of transcription. Typically, the two main control elements in eukaryotic transcription are a gene's promoter and its terminator. Terminators play an important role in controlling mRNA half-life, which has an important influence on the enzyme output levels. Given their decisive role, native expression-enhancing terminators have been intensively characterized and synthetic terminators improving heterologous gene expression have been developed [24,188,189]. Promoters on the other hand also have a very large impact on gene expression levels and are as such one of the most important parts of the yeast synthetic biology toolbox [190]. A select group of native yeast promoters is broadly used [191–193], typically representing constitutive and inducible promoters. Commonly used constitutive promoters ensuring gene expression in all conditions are the *TEF1*, *TDH3*, *CYC1* and *ADH1* promoters [159]. Inducible promoters on the other hand allow controllable expression and are activated when desired. Regularly used are the *GAL* and *CUP1* promoters, induced by galactose and copper respectively [190]. In general, constitutive promoters are preferred due to some inherent disadvantages of inducible promoters, such as lag time after induction, leaky expression and potential high inducer costs or inducer toxicity.

The structure of a eukaryotic promoter is well studied and they are generally divided in a core promoter element and upstream regulatory elements (Figure 3.1) [194]. The core promoter is the regulatory sequence to which RNA polymerase II binds and where transcription is started [195–199]. Therefore, it is seen as a major determinant of gene expression in yeast [196]. The length of the core promoter is typically around 100 – 200 bp and contains a nucleosome free region to enhance access of the pre-initiation complex (PIC) to the DNA. The PIC binds to the consensus TATA box, or a weaker TATA-like sequence differing up to 2 bp with the consensus, and scans the core promoter in search for a suitable transcription start site (TSS) [196]. Though some TSS consensus sequences have been suggested, *i.e.* RRY<u>R</u>R, TC<u>R</u>A, Y<u>A</u>WR and A(A_{rich})$_5$NY<u>A</u>WNN(A_{rich})$_6$, to date no fixed TSS

sequence in yeast has been agreed on [195]. Generally, transcription is initiated 40 to 120 bp further downstream of the TATA box or the TATA-like sequence in case of TATA-less promoters [195]. Core promoter activity was also observed to be higher with a pyrimidine rich scanning region and an adenine enriched initiation region [196]. Upstream regulatory elements are placed in front of the core promoter and typically contain one or more transcription factor binding sites (TFBSs). TFBSs are typically distributed between 50 and 150 bp upstream of the TSS and showed an enriched peak at 115 bp [200]. These cis-acting regulatory DNA stretches recruit transcription factors interacting with one another and with the basal transcriptional systems to regulate promoter activity [194]. As such, transcription factors can be repressors or activators of transcription and bind either to their respective upstream repressive sequence (URS) or upstream activating sequence (UAS) [201]. Promoter engineering strategies by both modulating the core promoter and upstream regulatory DNA elements are thus very effective ways to alter a gene's expression and hence to balance biosynthetic pathways [110].

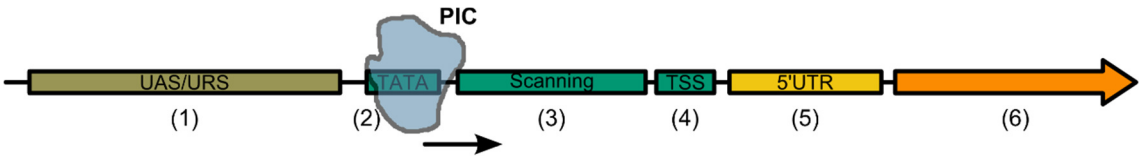

Figure 3.1: Schematic representation of the structure of a yeast promoter divided in upstream activating or repressive sequences (UAS/URS, 1) and the core promoter region (2, 3 and 4). The pre-initiation complex (PIC) containing RNA polymerase II is recruited to the TATA box or a TATA-like sequence (2) and scans (3) the core promoter to a suitable transcription start site (TSS, 4). Other depicted elements are the 5' untranslated region (5'UTR, 5) and the coding sequence (6).

Many approaches for promoter engineering in yeast have already been developed [190,194]. Well-known pioneering examples are error-prone PCR, hybrid promoter engineering and nucleosome affinity modulation. Error-prone PCR involves the introduction of random mutations in an existing promoter sequence. This strategy led for example to a 15-fold range promoter library of the popular *TEF1* promoter [22,202]. In the hybrid approach, the core promoter and UASs are seen as modular building blocks where a core promoter can be combined with one or multiple (different) UASs to alter total promoter activity [110,157]. Lastly, specific mutations suggested by predictive models decreased the nucleosome affinity in promoter sequences and resulted in a more open promoter structure, improving the access of transcription factors and thus enhancing transcription [112]. Despite these promising

reports though, no standardized or generally used method to quickly assemble yeast promoters with a broad range of different strengths, *e.g.* by using different UASs, exists. Currently such methods use classical restriction – ligation, which makes iterative assembly unavoidable if multiple UAS elements have to be integrated (*i.e.* repeated restriction – ligation reactions have to be performed) [110]. Moreover, the synthetic biology field of *S. cerevisiae* is still hampered by the big length of its promoters. Indeed, native promoters in yeast typically span ranges of hundreds of nucleotides, needed for the recruitment of the large RNA polymerase II. One way to solve this bottleneck could be the use of short, viral promoters such as the bacteriophage T7 promoter having a length of around 20 bp. Although the *in vivo* production of RNA transcripts with this orthogonal system has been demonstrated in *S. cerevisiae* [203,204], it has some inherent disadvantages like the requirement of a heterologous expressed T7 RNA polymerase and the inability of translating T7 transcripts due to the lack of a 7-methylguanosine cap which is necessary to initiate translation. Therefore, the construction of short promoters that interact with the native yeast RNA polymerase II is preferred. Currently, the library with the shortest synthetic promoters reported is the one of Redden *et al.* [111], with a length of around 100 bp. Together with the fact that every gene in eukaryotes needs its own promoter and terminator, the construction of large biosynthetic pathways in yeast quickly becomes a laborious task. In addition, most available and characterized yeast promoters today are based on native sequences [191,192]. This could promote homologous recombination between the different regulatory elements within the heterologous pathway and with the genome, leading to strain instability.

These hurdles could be tackled by the design of short yeast promoters with a range of different strengths comparable to those of broadly used native constitutive promoters [192]. Preferably, they should be less than 100 nucleotides in length for easy incorporation in primers, enabling fast transcription unit (TU) construction via PCR. As such, this study describes the development and characterization of a set of short yeast core promoters. We first identified by way of truncation the minimal length needed for transcription initiation of the well-characterized *TEF1* core promoter. Next, a library of semi-synthetic yeast core promoters (< 70 bp) was constructed by randomization of this *TEF1* minimal core promoter. Finally, we developed and evaluated a standardized yeast UAS Golden Gate method (yeast UAS Golden Gate, yUGG), for the fast and random assembly of a yeast promoter library, as no plug-and-play technique has yet been described for the fast and

random development of yeast promoters with varying strengths, based on existing core and upstream elements. The developed library was based on the native *TEF1* core promoter and different selected UASs, given the demonstrated potential of the hybrid promoter engineering approach [110].

3.3 MATERIAL AND METHODS

Unless otherwise stated, all products were purchased from Sigma-Aldrich (Diegem, Belgium), all fragments were PCR purified using the innuPREP PCRpure Kit (Analytik Jena AG, Jena, Germany), Circular Polymerase Extension Cloning (CPEC) [11] was used for the assembly of all plasmids and plasmid extraction was performed with the innuPREP Plasmid Mini Kit (Analytik Jena AG).

3.3.1 Strains and media

S. cerevisiae SY992 (*Matα, ura3Δ0, his3Δ1, leu2Δ0, trp1-63, ade2Δ0, lys2Δ0, ADE8*, Euroscarf, University of Frankfurt, Germany [205]) was used as yeast expression host. All yeast strains derived from this strain are listed in Supplementary Table S.1.1. Yeast cultures were grown in synthetic defined (SD) medium consisting of 0.67% YNB without amino acids, 2% glucose (Cargill, Sas van Gent, The Netherlands) and selective amino acid supplement mixture without uracil (CSM – URA, MP Biomedicals, Brussel, Belgium). To solidify media, 2% Agar Noble (Difco, Erembodegem, Belgium) was added.

Transformax™ EC100™ Electrocompetent *E. coli* (Lucigen, Halle-Zoersel, Belgium) was used for cloning procedures and for maintaining plasmids. *E. coli* strains were cultured in Lysogeny Broth (LB) consisting of 1% tryptone-peptone (Difco), 0.5% yeast extract (Difco), 1% sodium chloride (VWR, Leuven, Belgium) and 100 µg/ml ampicillin or 25 µg/ml chloramphenicol dependent on the selection marker. For the selection of *E. coli* strains after Golden Gate, sucrose medium without salt existing of 1% tryptone-peptone (Difco), 0.5% yeast extract (Difco) and 5% sucrose was used. For solid growth medium, 1% agar (Biokar diagnostics, Pantin Cedex, France) was added.

3.3.2 Plasmid construction

For the evaluation of the truncated *TEF1* library, five reference vectors with a yECitrine transcription unit (pKT140, Euroscarf [206]) under the transcriptional control of the *TEF1* [159], *ADH1* [159], *CYC1* [159], *PGK1* [191] or *TDH3* [159] promoter and the *ADH1* terminator [206] were constructed. These TUs were assembled on an in-house low copy yeast expression backbone consisting of a *CEN6/ARS4* origin of replication (ori) and a *URA3* auxotrophic marker (p2a backbone), resulting in the vectors pRef-pTEF1 (Supplementary Figure S.1.1), pRef-pADH1, pRef-pCYC1, pRef-pPGK1 and pRef-pTDH3 (Supplementary Table S.1.2). All

promoter and terminator sequences were picked up from the *S. cerevisiae* SY992 genome with PrimeSTAR HS DNA polymerase (Takara, Westburg, Leusden, The Netherlands). Vector pRef-pTEF1 was further used as template for the construction of the truncated *TEF1* core promoter library plasmids (Figure 3.2). The core promoter sequence specified by Blazeck *et al.* [110] was shortened by ca. 20 bp per time through primers containing overlap sequences for CPEC (Integrated DNA Technologies, Leuven, Belgium, and Supplementary Table S.1.3 and Figure S.1.1). More specifically, the p2a backbone was split by two primers (o_BBsplit_fw and o_BBsplit_rv, Supplementary Table S.1.3). Two pieces CPEC, consisting of a part PCR-amplified by the forward core promoter primer (o_UAScpTEF_1 to 9 or o_cpTEF_1 to 9) and o_BBsplit_rv and a fixed backbone part PCR-amplified by o_BBsplit_fw and o_BBUAScpTEF or o_BBcpTEF, was performed leading to respectively p_UAS-cpTEF_1 to 9 and p_cpTEF_1 to 9.

Plasmid p_cpTEF_6, containing the 69 bp long minimal *TEF1* core promoter, was used as template for the construction of four *TEF1* core promoter libraries. Four oligonucleotides each containing 18 degeneracies (IDT, Supplementary Table S.1.3) and covering together the whole length of the minimal core promoter were ordered (Figure 3.3). After plasmid assembly, the distribution of degeneracies (ca. 25% of each nucleotide) was confirmed by sequencing (EZ-Seq, Macrogen, Amsterdam, The Netherlands).

For the construction of a random library with one or multiple UASs, Golden Gate (GG) was used as assembly method [9]. Therefore, the different UAS parts were flanked by inward-facing AarI sites and assembled in GG carrier vectors (pJET backbone, ThermoFisher Scientific, Aalst, Belgium). Two types of GG carrier vectors were constructed: (i) carrier vector type-M resulting in sticky ends for multiple integration events in the destination vector and (ii) carrier vector type-S resulting in sticky ends for a single integration event (Figure 3.6A). The GG destination vector contained outward-facing AarI sites flanking a *SacB* gene which is replaced in correctly assembled expression vectors by the UASs and enables screening of correct *E. coli* colonies on sucrose medium without salt [207] (Supplementary Figure S.1.2). The destination vector contained furthermore the yECitrine TU under control of the *TEF1* core promoter specified by Blazeck *et al.* [110] and its native 5'UTR, the *ADH1* terminator, and maintenance elements: a chloramphenicol resistance marker and pUC ori for *E. coli*, and a *CEN6/ARS4* ori and *URA3* marker for yeast PCR amplified from the p2a backbone (Figure 3.6A).

For the Golden Gate assembly of the UAS library, 100 ng GG destination vector with a 50-fold molar excess of carrier vector type-M and a 10-fold molar excess of carrier vector type-S were mixed in a one-pot GG reaction (total volume of 20 µl). The reaction mixture consisted of 2 µl 10x T4 ligase buffer (ThermoFisher), 0.4 µl 50x oligonucleotides for AarI (ThermoFisher), 2 U of AarI (ThermoFisher) and 60 Weiss units of T4 DNA ligase (ThermoFisher). The restriction-ligation was carried out in a thermocycler during 50 cycles of 2 min at 37°C and 3 min at 16°C. The reaction was stopped by two final steps of 10 min at respectively 50°C and 80°C. Afterwards, 2 µl of GG reaction mixture was electroporated in Transformax™ EC100™ Electrocompetent *E. coli* (Lucigen) and plated on salt-lacking sucrose plates containing 25 µg/ml chloramphenicol. For every experiment, 16 colonies were evaluated by colony PCR and plasmids were verified by sequencing (EZ-Seq, Macrogen).

All plasmids and libraries were transformed in yeast SY992 via the lithium-acetate method [208]. After transformation, strains were selected on SD CSM – URA plates and confirmed by yeast colony PCR using *Taq* DNA polymerase (NEB, Bioké, Leiden, The Netherlands). For library evaluation, single colonies were randomly picked from agar plate and grown in 96-well microtiter plates (MTPs).

An overview of all plasmids used and constructed in this study can be found in Supplementary Table S.1.2.

3.3.3 Fluorescence and absorbance measurements

Fluorescence was used here as a measure of protein levels. Fluorescent proteins are optimized and generally known to fold very well in bacteria and yeasts. As such, it is supposed that these proteins are synthesized in a fully active form and that higher fluorescence levels correspond to more production of the protein. Fluorescence as measure of protein abundance is also widely accepted in the field of synthetic biology [23,30,110,111,181,209].

Four biological replicates were inoculated from agar plate in sterile 96-well flat-bottomed, black microtiter plates (Greiner Bio-One, Vilvoorde, Belgium) enclosed by a Breathe-Easy® sealing membrane (Sigma-Aldrich) containing 150 µl selective SD CSM – URA medium. These plates were incubated on a Compact Digital Microplate Shaker (ThermoFisher Scientific, 3 mm orbit) at 800 rpm and 30°C for 24h. Subsequently, these pre-cultures were diluted 150 times in 150 µl fresh selective SD CSM – URA medium and grown in sterile

polystyrene black µclear flat-bottomed 96 well plates (Greiner Bio-One) for evaluation. Except for the randomized *TEF1* core promoter library, the µclear flat-bottomed plate cultures were evaluated in continuous growth experiments performed in a TECAN Infinite® 200 PRO MTP reader (Tecan). Optical density (OD, 600 nm) and fluorescence (FP, excitation and emission of yECitrine, 502 nm and 532 nm, respectively) were measured every 15 min for 50 hours at 30°C (orbital shaking at 2 mm orbit). For every strain, the endpoint OD was determined as the OD value after which three descending OD values were observed. This endpoint OD with its corresponding FP value were used for further data analysis. For the evaluation of the randomized *TEF1* core promoter library, an endpoint OD and FP measurement was taken after 26h of growth (stationary phase) at 30°C while shaking at 800 rpm (Compact Digital Microplate Shaker, 3 mm orbit).

For analysis of fluorescence measurements, two types of controls were included in every single MTP. A medium blank (*i.e.* SD CSM – URA medium) was used for the correction of background absorbance of the medium (OD_{bg}). sRef-bl lacking fluorescent protein expression and containing p2a_empty was used to correct for the background fluorescence of yeast (FP_{bg}). Fluorescence corrected for OD was used as measure for fluorescent protein expression and calculated as follows:

$$\left(\frac{FP}{OD}\right)_{corrected} = \frac{FP - FP_{bg}}{OD - OD_{bg}} \tag{3.1}$$

The relative fluorescence was defined as follows:

$$Relative\ fluorescence\ (\%) = \frac{\left(\frac{FP}{OD}\right)_{corrected}}{\left(\frac{FP}{OD}\right)_{corrected,\ ref}} \times 100 \tag{3.2}$$

3.3.4 Data analysis

All calculations were performed in Python using the Python Data Analysis Library (Pandas). Error bars represent the standard error of the mean (n = 4). Pairwise comparisons between different strains were done by a two-sided T-test using the scipy.stats package in Python. In all cases, a significance level of 0.05 was applied.

3.4 RESULTS AND DISCUSSION

3.4.1 The minimal TEF1 core promoter

For the development of minimal yeast core promoters, a yECitrine based fluorescent reporter system to evaluate the altered promoter influence on gene expression was constructed. As a starting point, the *Saccharomyces cerevisiae TEF1* promoter, which is well described in literature [22,202] and is used a lot in synthetic biology approaches in yeast [15,20,192,210–212], was used. Based on the categorization of Blazeck *et al.* [110], the *TEF1* promoter is divided in an upstream activating sequence (UAS) and core promoter, *i.e.* a 203 bp long UAS and a 176 bp long core promoter. Furthermore, *TEF1* has a 5'UTR of 33 bp. Since this promoter has no fixed TATA box and no defined transcription start site is described, the minimal *TEF1* core promoter was first determined by truncation of the core promoter in steps of ca. 20 bp toward its 5'UTR. In addition, the effect of the native UAS on minimal core promoter activity was investigated by developing two sets of truncated promoters, one with and one without the UAS. As such, two libraries of nine promoters with different lengths were constructed (Figure 3.2A).

Truncating the *TEF1* core promoter led to an overall decrease in protein expression for both libraries (Figure 3.2B), which was expected since the large RNA polymerase II complex needs a long stretch of DNA for binding and stabilization [213]. However, the truncation of cpTEF_3 to cpTEF_4, which resulted in the complete deletion of a poly-dT stretch, caused an increase in promoter activity. It is reported that such long stretches of consecutive similar nucleotides have an influence on nucleosome affinity. Especially the complement stretch, *i.e.* a poly-dA tract, disfavors nucleosome formation and thus promotes binding of regulatory promoter elements which enhances total promoter activity [214–217]. In addition, poly-dA:dT tracts drive nucleosome positioning in the promoter by creating boundaries against which nucleosomes are located [218]. It was indeed observed that nucleosome organization was dramatically changed when a poly-dA:dT tract with its upstream promoter region was deleted [218]. In general, poly-dA:dT tracts are very influential parts crucial for accurate nucleosome organization and thus playing a determining role in the structure of a yeast promoter. This makes them also interesting targets to modify transcriptional regulation of gene expression. Hence, we suggest that this poly-dT removal in the *TEF1* core promoter could be a major cause of the sudden increase in yECitrine fluorescence. It is also noteworthy that the truncated library led to very short promoters

with activities higher than native long and weak yeast promoters. For example, cpTEF_5 existing of only 90 bp was 1.5-fold stronger than the 1500 bp long *ADH1* promoter. This confirms (minimal) core promoters as key determinants of gene expression levels.

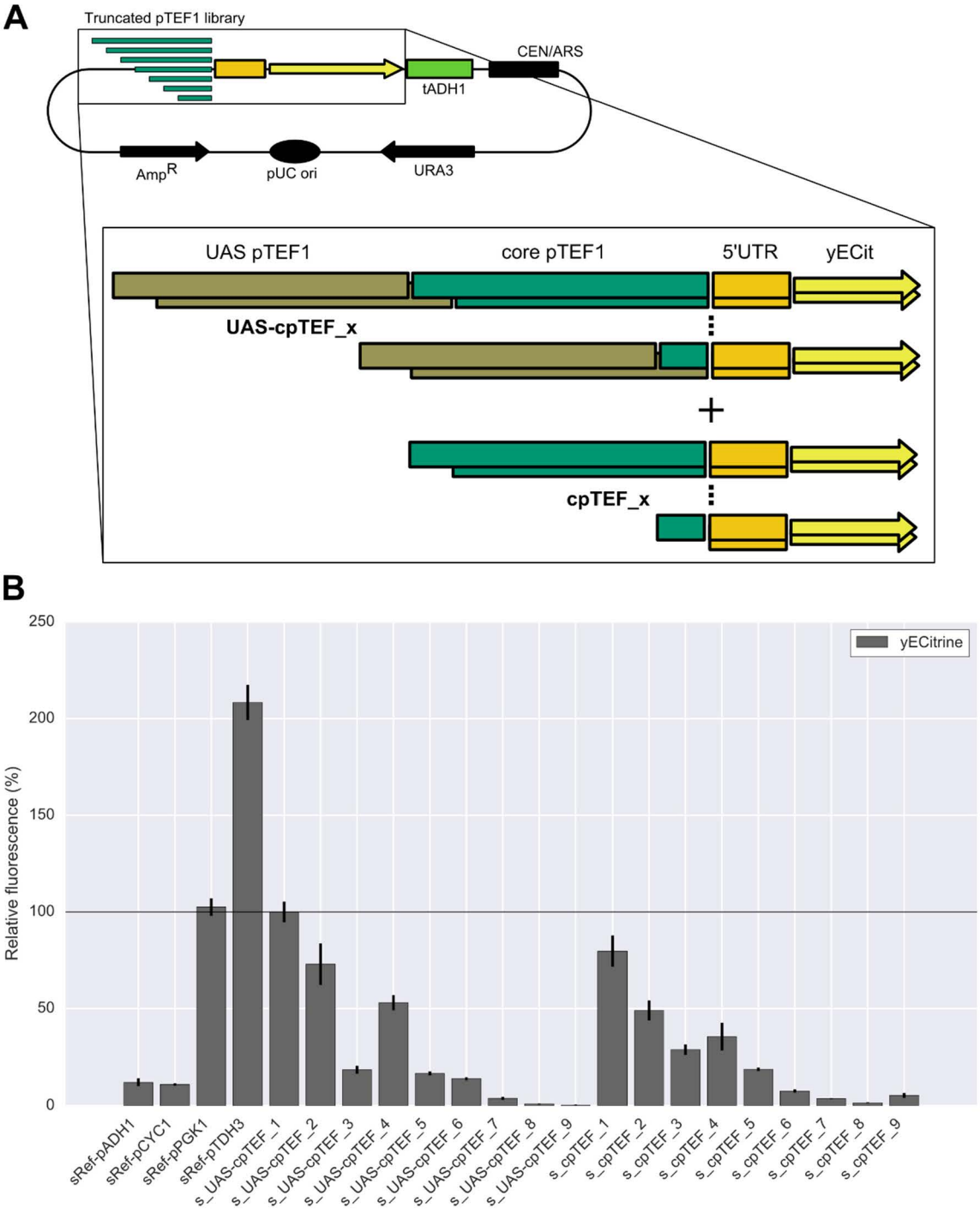

Figure 3.2: Truncated *TEF1* core promoter library. (A) Schematic overview of the *TEF1* core promoter libraries with (UAS-cpTEF_x) and without (cpTEF_x) the *TEF1* UAS; x varies from 1 to 9. The sequences of the truncated *TEF1* core promoters and the *TEF1* UAS are given in

Supplementary Table S.1.4 and Table S.1.5, respectively. (B) yECitrine fluorescence obtained with the truncated *TEF1* core promoter libraries and four reference promoters. The values are given relative to s_UAS-cpTEF_1 (horizontal line) representing the native *TEF1* promoter. Error bars represent the standard error of the mean (n = 4, biological repeats).

Pairwise comparisons of the UAS-cpTEF_x and cpTEF_x strains showed only a significant positive influence of the UAS on protein expression for strain s_UAS-cpTEF_6 (p = 6.73E-4) and a significant negative effect for strains s_UAS-cpTEF_3 and s_UAS-cpTEF_9 (p = 0.013 and p = 0.008, respectively). This was somewhat surprising since the addition of an UAS_{TEF1} in front of a truncated *LEU* promoter [110] significantly increased expression levels. However, when taking in mind that the p-values of strains 1, 2 and 4 were very close to the significance level ($p \approx 0.05$) and that others report similar results to ours in that respect that an extra *CIT1* or *CLB2* UAS in front of some synthetic core promoter elements also did not improve protein expression [111], our results are in line with earlier observations. For the shortest core promoter elements, UAS_{TEF1} does not influence expression levels, presumably by the inability of RNA polymerase II to bind, leading to complete failure of proper transcription initiation.

Altogether, since cpTEF_6 is the shortest core promoter giving rise to detectable transcription (*i.e.* ca. 0.66-fold lower than the weak *ADH1* and *CYC1* promoters) and showing significant activation by its UAS, this core promoter element was chosen for random library construction.

3.4.2 Random TEF1 core promoter library

For the construction of a range of short core promoters in *S. cerevisiae*, a randomization approach with degenerated oligo's spanning the 69 bp long core promoter cpTEF_6 was used (Figure 3.3). Both to allow us to capture positional effects of the randomization and to sample significant quantities of the created variance, cpTEF_6 was divided into four DNA stretches of 18 base pairs. One stretch per library was degenerated, whereas the others were kept constant. To have for every library the same space of possibilities, cpTEF_6-libD was extended with the first three base pairs of the *TEF1* 5'UTR. Evaluation of these four libraries could give an idea of the regions in cpTEF_6 that have a high influence on protein expression and are interesting for further analysis.

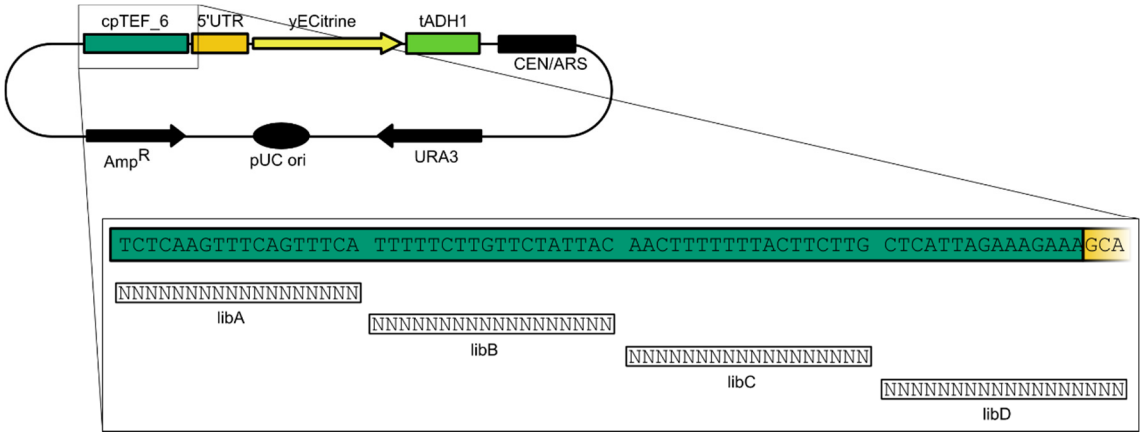

Figure 3.3: Strategy used for the construction of a minimal *TEF1* core promoter library. Promoter cpTEF_6 was divided into four equal DNA tracts where for every library 18 bp were randomized while the rest of the sequence remained unchanged. Library cpTEF_6-libD randomized also the first three base pairs of the *TEF1* 5'UTR.

From each library 94 colonies were randomly picked and evaluated for fluorescence (*i.e.* yECitrine). sRef-bl and s_cpTEF_6 were taken along as references in every MTP. Histograms of the four libraries revealed a small shift toward higher fluorescence levels for libraries cpTEF_6-libA and D (Figure 3.4, Supplementary Figure S.1.3 to Figure S.1.6, p-values when comparing the means for library A and D to library C and B at least < 0.01). Both library A and D gave thus more rise to promoters with higher strengths than the native cpTEF_6 compared to library B and C (12 and 13 versus 2 and 0). This indicates that both regions in the cpTEF_6 promoter are interesting to enhance transcription levels. Library D, varied immediately in front of the 5'UTR, including in the TSS, confirms that the region around the TSS is an important feature for initiation of mRNA transcription. It has indeed been described that the scanning of RNA polymerase II in search of a suitable TSS depends on its surrounding context [196,219–221]. Library A, varied at the 5' end of the cpTEF_6 promoter sequence, is positioned around 50 nucleotides away from the TSS. It has been reported that transcription in *S. cerevisiae* is started 40 to 120 bp downstream of the TATA box or a TATA-like sequence (in the case of pTEF1) and variation in this region alters core promoter activity [195]. As such, the PIC might bind and compose itself in this region, *i.e.* library A, of the core promoter. Libraries cpTEF_6-libB and C are suggested to form a scanning region of around 40 bp between the PIC binding place and the TSS and they seem to generate more promoters with lower activity. This is plausible as the native cpTEF_6 spacer region is

already enriched in T and C (~ 80%) and a T/C-rich scanning region was linked with higher expression levels [196].

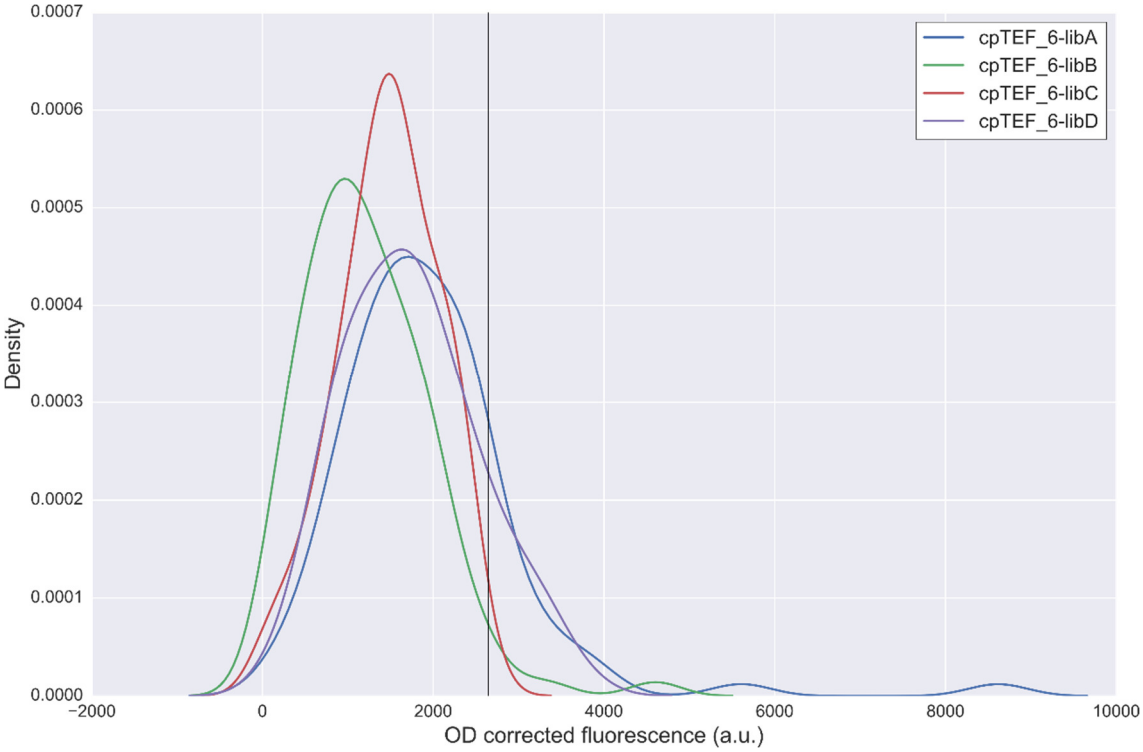

Figure 3.4: Histograms of the four randomized cpTEF_6 libraries after fluorescence analysis of 94 randomly chosen colonies. The vertical line represents the mean fluorescence (2645 ± 332.7 a.u., n = 4, biological repeats) of the native cpTEF_6 promoter.

As cpTEF_6-libA had a high potential to contain more high-expressing core promoters, 281 additional colonies were randomly selected and analyzed together with three biological repeats of s_cpTEF_6 and sRef-bl. Analysis of fluorescence levels revealed a similar distribution pattern as for the first 94 colonies analyzed (Supplementary Figure S.1.7). Again, some very high expressing core promoters were identified compared to the native sequence, a result which was not achieved by error-prone PCR of the whole *TEF1* promoter [202]. The four strongest and three weaker ones were chosen for further characterization. The promoter activity of this range of seven 69 bp long core promoters was compared to five commonly used native yeast promoters, *i.e.* pADH1, pCYC1, pTEF1, pPGK1 and pTDH3 (Figure 3.5). Interestingly, the results show that we obtained promoters (s_cpTEF_6-F and G) having a 2.0 to 4.0-fold higher strength than the native cpTEF_6 minimal core promoter (s_cpTEF_6) we started from. In addition, while using a sequence of only 69 bp, s_cpTEF_6-

G and s_cpTEF_6-D, E and F respectively led to double and equal transcript levels compared to the weak *ADH1* (1500 bp) and *CYC1* (287 bp) promoters. Furthermore, the strong native pTEF1 is only 2.8-fold stronger than the short cpTEF_6-G, but the latter has a reduction of 82% in sequence length. As such, the reported set of promoters shows that an adequate expression range (4.0-fold) can be achieved using core promoter elements smaller than 70 bp.

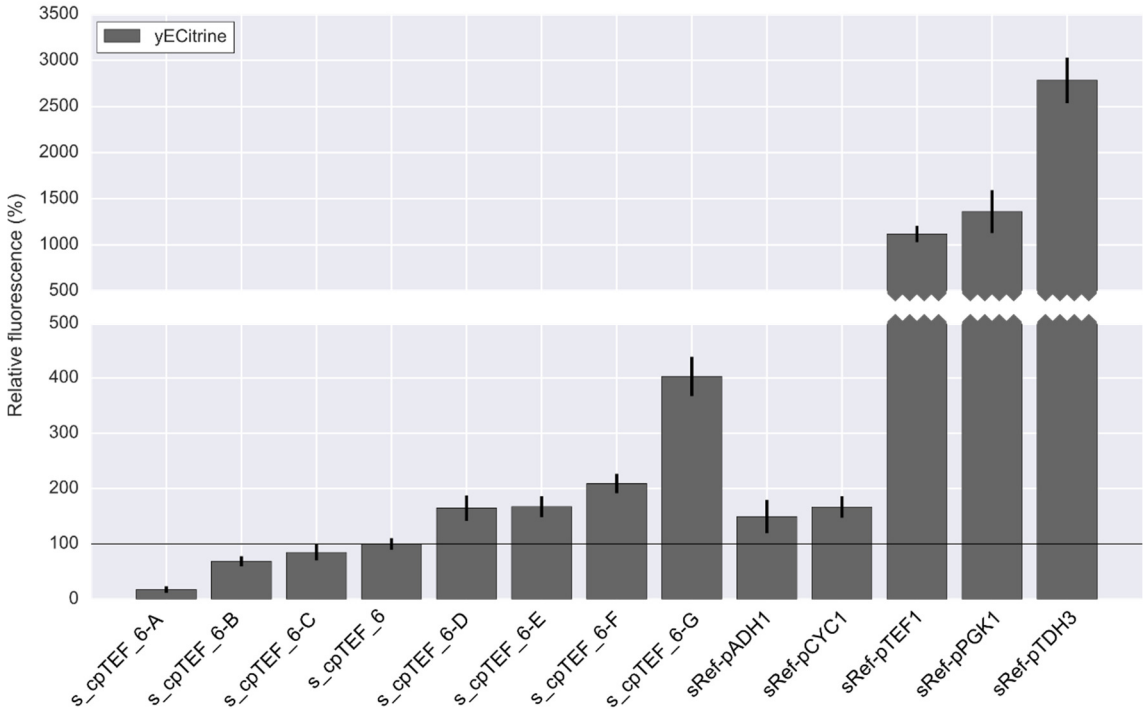

Figure 3.5: Characterization of seven selected promoters from library cpTEF_6-libA. Protein expression levels were normalized against the native cpTEF_6 promoter (horizontal line). Error bars represent the standard error of the mean (n = 4, biological repeats). All strains are listed in Supplementary Table S.1.1.

Although a set of nine minimal core promoters has been described before [111], the set described here consists of short promoters that are stronger than for example the native *CYC1* promoter without the need of UASs. In earlier observations an extra UAS had to be added to the core promoters to reach higher expression levels than the native *CYC1* promoter, leading to promoters over the 100 bp in length [111]. Still, the obtained promoter library of short core promoters has the disadvantage to form a less broader expression range compared to the earlier reported 15-fold *TEF1* promoter library [202]. The main difference to our approach is that Nevoigt and coworkers varied the whole 412 bp long *TEF1*

promoter (*i.e.* core promoter inclusive its UAS). As such, not only the sequence to bind RNA polymerase II is altered, but also the sequence to recruit transcription factors which could consequently lead to extra variation in gene expression. Therefore, in view to broaden our expression range, a Golden Gate method to randomly introduce UAS sequences in front of a yeast core promoter was developed.

3.4.3 yUGG, a method for one step, random assembly of an UAS library

In a next step, we applied hybrid promoter engineering to obtain stronger promoters and broader expression ranges. Hybrid promoter engineering, with UAS elements serving as modular amplifiers in front of a core promoter, has proven to be an effective technique for the enhancement of gene expression levels in yeast [110,157]. Unfortunately though, no method is described for the one step and random assembly of multiple UAS elements in front of a core promoter. Especially with our set of short core promoters such a random approach could further enlarge the range of expression strengths. As such, a Golden Gate inspired method, yeast UAS Golden Gate (yUGG), was developed, allowing the construction of a random UAS library in front of a core promoter. Three previously identified UAS elements from constitutive promoters were selected: UAS_{CIT1} of the mitochondrial citrate synthase *CIT1* gene, UAS_{CLB2} of the mitotic cyclin *CLB2* gene and UAS_{TEF1} from the translational elongation factor EF-1α (Supplementary Table S.1.5) [110]. The native cpTEF_1 promoter, the strongest core promoter of the truncated library, was chosen as proof of principle to assess the potential of the yUGG method. The yUGG method is designed with AarI, a rare cutter type IIs restriction enzyme with a non-palindromic heptanucleotide recognition site producing 4 bp sticky ends [222]. For each UAS, two types of GG carrier vectors were constructed; one with equal 5' and 3' sticky ends for multiple integration (type-M) and one with different 5' and 3' sticky ends for single integration (type-S). Together with a destination vector containing a different 5' and 3' sticky end, this design makes it possible to introduce one or multiple UASs without self-closure of the destination vector (Figure 3.6A). In addition, the set of six carrier vectors makes it possible to just use one type of UAS or play with different combinations of UASs, *e.g.* all three UASs or only UAS_{CIT1} and UAS_{CLB2}, *etc*.

First, assembly efficiency was optimized by varying the ratio between the destination vector and carrier vectors. Specifically, 1:1:5, 1:5:25, 1:10:50, 1:20:100 and 1:50:250 molar ratios of respectively the destination vector, carrier vectors type-S and carrier vectors type-M

were evaluated. The efficiency of obtaining correctly assembled vectors, as well as of obtaining a variation of vectors, was determined via colony PCR on 16 randomly selected colonies (Supplementary Figure S.1.8 shows as an example the difference in length of various built-in UASs that can be observed via colony PCR). Efficiency was the highest when using the 1:10:50 ratio. Next, three one-pot GG reactions were performed to determine up to how many copies of a single UAS type can be introduced efficiently (*e.g.* just pU-CLB2-S and pU-CLB2-M). Up to four UAS elements could be incorporated in one step (Supplementary Table S.1.2) with a high prevalence of one or two copies of UASs. Keeping in mind the very efficient homologous recombination machinery of *S. cerevisiae*, incorporation of more than four similar UAS elements would anyhow not be recommendable, since this could lead to recombination-based promoter instability. In a final step, one GG assembly using all three UAS elements (*i.e.* all six carrier vectors) to combine multiple diverse UASs was carried out. Although based on colony PCR it seemed that efficient diversification was obtained, sequencing revealed that combining the different UASs was not as efficient as expected. Only two plasmids out of 10 sequenced contained distinctive UAS elements (Supplementary Table S.1.2), while the others had either only one UAS or a combination of equal UASs assembled. Screening efforts hence become an obstacle, as a lot of combination events are possible and the UAS building blocks are similar in length. Though this is not a huge issue if only a range of different promoter strengths needs to be obtained. On the other hand, if a specific order and type of distinct UAS elements is desired, it would be better to expand our method for sequential assembly purposes such as PODAC, which enables iterative GG with a single restriction enzyme [131].

The effect of the different UAS elements on transcription initiation was evaluated by measuring yECitrine fluorescence (Figure 3.6B). To start with, it is noteworthy to mention that fluorescence levels obtained with sRef-pTEF1, where expression is controlled by the natural pTEF1, and s_UAS_{TEF1}-1x, where expression is controlled by the yUGG based cpTEF_1 with 1 UAS, are the same. Furthermore, in contrast to earlier findings using a truncated *LEU* promoter [110], the addition of multiple UAS_{TEF1} elements did not significantly increase transcription levels. Similar results are obtained with UAS_{CLB2}: although the addition of 1 UAS_{CLB2} significantly increases fluorescence ($p = 0.029$), additional copies of UAS_{CLB2} do not lead to a significantly higher fluorescence. When comparing strains s_$UAS_{CIT1\text{-}CLB2}$ and s_$UAS_{CIT1\text{-}TEF1\text{-}CLB2}$, the extra UAS_{TEF1} also has no significant effect on fluorescence levels. On the other hand, tandem repeats of UAS_{CIT1} gradually and significantly increase

yECitrine expression (p-values < 9.2E-3). The introduction of three UAS_{CIT1} elements in front of cpTEF_1 led to a 2.5-fold stronger promoter than cpTEF_1 and a 2.0-fold increase compared to the native *TEF1* promoter, which is a strong yeast promoter. UAS_{CIT1} has been reported as more effective than UAS_{CLB2} in front of the native *TEF1* promoter [110], which could explain its excellent behavior in front of the short cpTEF_1 promoter. The single combination of all three UAS elements ($s\_UAS_{CIT1\text{-}TEF1\text{-}CLB2}$) led to equal expression levels of s_UAS_{TEF1}-3x (p = 0.10) and s_UAS_{CLB2}-3x (p = 0.53) which was, for the latter, also observed in front of the *TDH3* promoter [110]. Nevertheless, fluoresence levels were significantly lower compared to s_UAS_{CIT1}-3x (p = 5.0E-3).

To conclude, an increasing number of UAS elements can improve transcriptional activity, yet the degree of change in transcription levels is strongly dependent of the combination of a specific UAS type with a specific core promoter. This confirms an important interplay between UASs and core promoters, so both must be compatible for the enhancement of transcription [201,223]. As such, UAS elements can be used as modular building blocks to amplify transcription, but the magnitude of their effect is difficult to predict. Nevertheless, the aim of obtaining stronger promoters using hybrid promoter engineering was reached, as all hybrid promoters obtained are stronger than the cpTEF_1 promoter.

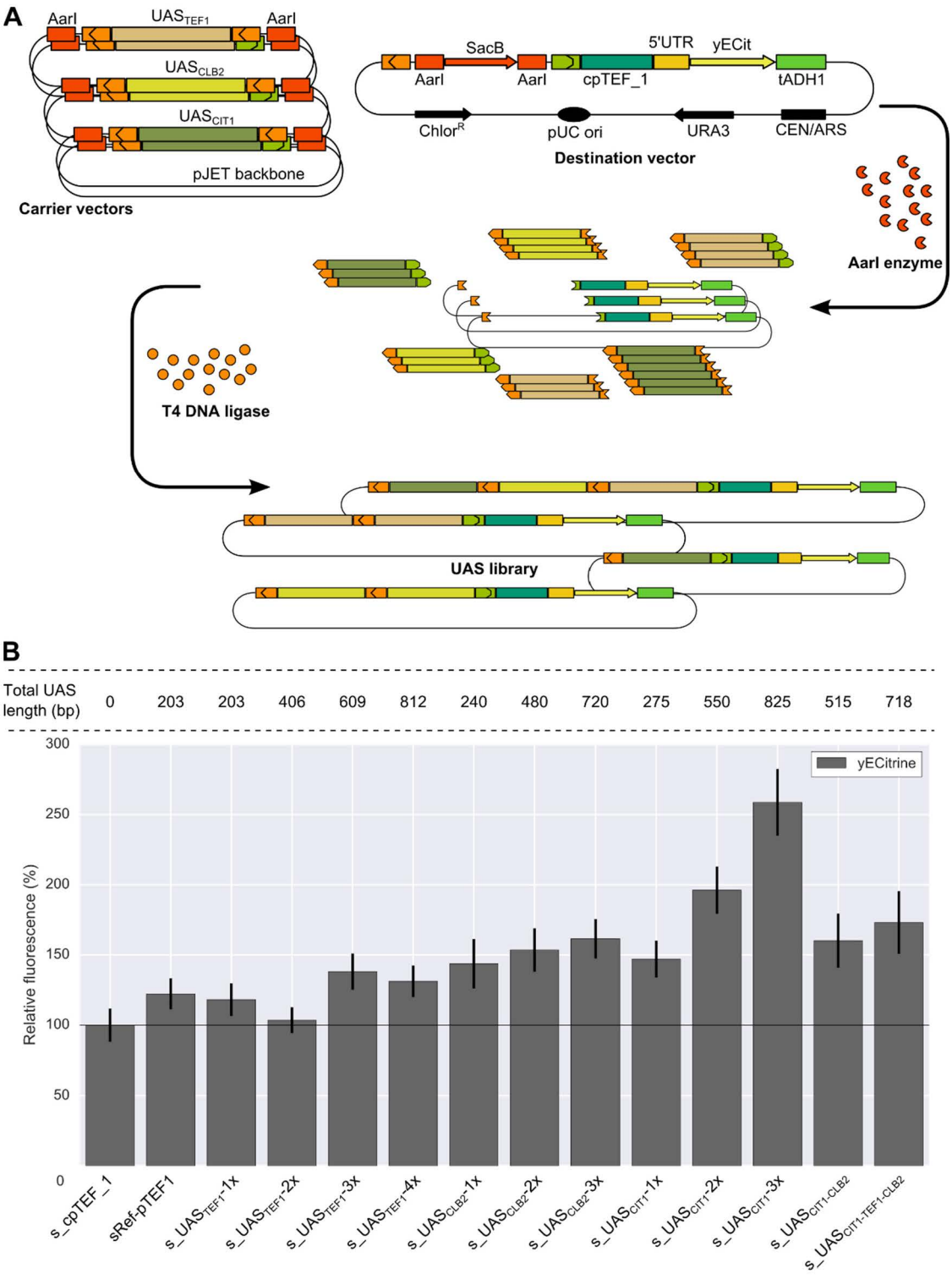

Figure 3.6: UAS based hybrid promoter engineering using Golden Gate. (A) Schematic overview of the one-pot Golden Gate assembly using AarI and T4 DNA ligase for the development of an UAS library. (B) Relative fluorescence against s_cpTEF_1 (horizontal line and carrying the native *TEF1* core promoter cpTEF_1) of the different obtained UAS library members. In addition, the total length of combined UASs in front of the core promoter is given. Error bars represent the standard error of the mean (n = 4, biological repeats).

3.5 CONCLUSION

Promoters play by far the most eminent role in transcriptional regulation of living cells and are as such indispensable for bio-engineering purposes. Especially with the ever increasing complexity of heterologous pathways built in microbes such as yeast, a further expansion of these regulatory parts remains necessary. In this study, the well-characterized *TEF1* promoter was truncated to elucidate its minimal core promoter and to enable the design of short functional yeast promoters. Six of the nine truncated promoters remained functional, even without the addition of an UAS and a 69 bp long *TEF1* minimal core promoter was determined. Randomization of this core promoter sequence revealed influential regions at the 5' and 3' ends, respectively suggesting a location of the PIC region and confirming the importance of the region around the TSS. This randomization approach led to short semi-synthetic yeast promoters with a length of 69 bp and more than twice as strong as the native *ADH1* and *CYC1* promoters. Especially for the reason that homologies of 20 to 30 bp are already sufficient to initiate homologous recombination in *S. cerevisiae*, an additional future perspective could be to check the genomic stability of these promoters. Nevertheless, homologies of 60 bp or more are needed for highly efficient homologous recombination which is not the case for our promoters having a native part of around 50 bp.

To obtain with these short promoters activities as strong as those of the strongest reported yeast promoters, *i.e.* to enlarge the activity range to a similar level as the activity range available for frequently used natural promoters, hybrid promoter engineering was applied. A standardized method, yUGG, for one-step UAS library construction was developed. Up to four UAS elements and three different UASs could be incorporated in one assembly reaction. In these experiments, fluorescence measurements showed the high potential of UAS_{CIT1} in front of a truncated *TEF1* promoter to enhance transcription. Altogether, these results demonstrate the possibility of short yeast promoter libraries, of which the expression range can be expanded with UASs, to function as full transcriptional regulators. In future work, our standardized assembly method should be expanded with our set of short core promoters (*e.g.* cpTEF_6-G) and extra UAS elements [111], for the fast development of collections of (short) yeast promoters having a broad variety of strengths, a critical requirement for metabolic engineering applications.

CHAPTER 4 TOWARD PREDICTABLE 5'UTRs IN *SACCHAROMYCES CEREVISIAE*: DEVELOPMENT OF A yUTR CALCULATOR

4.1 ABSTRACT 61
4.2 INTRODUCTION 62
4.3 MATERIAL AND METHODS 65
4.3.1 Strains and media 65
4.3.2 Plasmid construction 65
4.3.3 In vivo fluorescence measurements 66
4.3.4 Model feature quantification 67
4.3.5 Partial least squares (PLS) regression 68
4.3.6 Search algorithm for de novo design of 5'UTRs 69
4.3.7 Cultivation of p-coumaric acid production strains 69
4.3.8 Detection and quantification of p-coumaric acid 70
4.3.9 Data analysis 70
4.4 RESULTS AND DISCUSSION 71
4.4.1 Development of the yUTR calculator 71
4.4.2 The yUTR calculator compared to the Dvir model 75
4.4.3 Universal applicability of the yUTR calculator 77
4.4.4 Protein coding sequence influence and reverse engineering 79
4.4.5 Proof of concept: reliable p-coumaric acid production 80
4.5 CONCLUSION 83

Authors:

Thomas Decoene, Gert Peters, Sofie De Maeseneire and Marjan De Mey

This chapter has been published as:

Decoene, T., Peters, G., De Maeseneire, S., & De Mey, M. (2018). Toward predictable 5'UTRs in *Saccharomyces cerevisiae*: Development of a yUTR calculator. *ACS Synthetic Biology*. https://doi.org/10.1021/acssynbio.7b00366

Author contributions:

TD, GP, SDM and MDM were involved in the conception and design. TD, SDM and MDM drafted the manuscript. PLS model and yUTR calculator construction was performed by GP. TD performed the experiments, data analysis and interpretation of the results. All authors revised the manuscript critically.

Note:

The code used for the development of the model and the yUTR calculator is available on GitHub at https://github.com/DeMeylab/2018---yUTR-calculator.

4.1 ABSTRACT

Fine-tuning biosynthetic pathways is crucial for the development of economic feasible microbial cell factories. Therefore, the use of computational models able to predictably design regulatory sequences for pathway engineering proves to be a valuable tool, especially for modifying genes at the translational level. In this study we developed a computational approach for the *de novo* design of 5'-untranslated regions (5'UTRs) in *Saccharomyces cerevisiae* with a predictive outcome on translation initiation rate. Based on existing data, a partial least square (PLS) regression model was trained and showed good performance on predicting protein abundances of an independent test set. This model was further used for the construction of a 'yUTR calculator' that can design 5'UTR sequences with a diverse range of desired translation efficiencies. The predictive power of our yUTR calculator was confirmed *in vivo* by different representative case studies. As such, these results show the great potential of data driven approaches for reliable pathway engineering in *S. cerevisiae*.

4.2 INTRODUCTION

The ability to precisely control fluxes through biosynthetic pathways in living cells is a fundamental requirement for the fast development of new microbial cell factories. Pathway optimization tools typically consider three control levels, *i.e.* transcription, translation and post-translation. Well-known efforts in *S. cerevisiae* at the transcriptional level are (synthetic) promoter and terminator libraries, synthetic transcription factors and modification of transcription factor binding sites (TFBS) [22,24,111,188,190,202]. Also in yeast, modifications at the post-translational level, using different isomers or replacing (single) amino acids, have led to modified and improved enzyme characteristics [64,212,224].

To tune a gene's translation, *i.e.* to alter and predict protein expression levels, varying the translation initiation rate has been proven valuable. Especially for *E. coli* several methods that can predict protein levels are available, such as the design of Shine-Dalgarno sequences that lead to protein expression levels within a desired target range. Eminent examples are the RBS Calculator [30,225] and EMOPEC [181]. With the latter for instance, 91% of the generated sequences had protein levels within twofold of the aimed target level [181]. Despite these successes in prokaryotes, less progress has been made in predicting the translation initiation rates for the regulation of protein levels in eukaryotes.

One of the main hurdles is the complexity of the eukaryotic translational regulation machinery, shortly described here but extensively reviewed by Hinnebusch *et al.* [226,227]. The translation process exists of four steps, *i.e.* initiation, elongation, termination and ribosome recycling. Herein, the initiation step is seen as the most complicated step. To initiate translation on an mRNA in *S. cerevisiae*, the 40S and 60S ribosomal subunits together with the methionyl-transfer RNA (Met-tRNA$_i^{Met}$) and 11 translation initiation factors are required, compared to only three translation factors in bacteria [227]. An important first step in this initiation is the formation of the 43S pre-initiation complex (PIC) consisting of the GTP-bound eukaryotic initiation factor 2 (eIF2), Met-tRNA$_i^{Met}$, the 40S subunit and four other initiation factors. The 43S PIC, together with the eIF4 family of initiation factors, binds the mRNA at the 7-methylguanosine (m^7G) cap at the 5' end of the mRNA and forms a 48S PIC. This complex scans the 5' untranslated region (5'UTR) of the mRNA in the 3' direction toward a suitable AUG start codon (Figure 4.1). Once the start codon is found, the 60S subunit joins the 48S PIC leading to the formation of an 80S initiation complex (IC), ready now to start the synthesis of the protein (*i.e.* protein elongation). Since the 48S PIC 'moves'

over the 5'UTR in search for a suitable start codon, it is clear that the 5' UTR plays a crucial role in the translation initiation process [228] and thus in protein expression. Therefore, engineering 5'UTR sequences that have a predictive impact on protein levels could be a very efficient way for the fine-tuning of protein expression in eukaryotes and more specifically *S. cerevisiae*. However, there are some inherent properties of eukaryotic 5'UTRs (Figure 4.1) that make them more complex than prokaryotic ones and consequently complicate the engineering process.

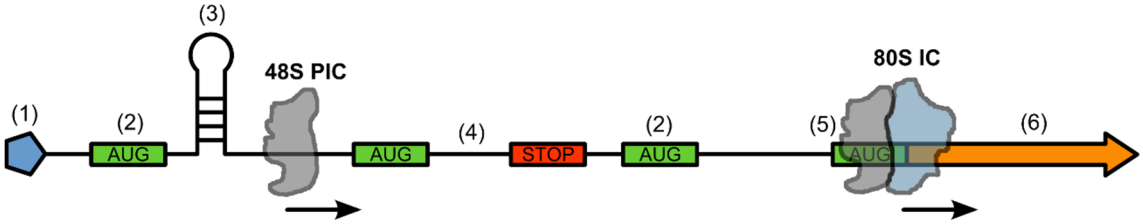

Figure 4.1: Schematic representation of the 48S pre-initiation complex scanning a eukaryotic 5'UTR and the formation of the 80S initiation complex at the correct start codon. In addition, possible elements interfering with the initiation of translation are depicted: (1) m^7G – cap. (2) Upstream AUG (uAUG). (3) Secondary structure. (4) Upstream open reading frame (uORF). (5) Main or authentic start codon. (6) Coding sequence. Abbreviations: PIC: pre-initiation complex; IC: initiation complex.

In contrast to prokaryotes where 5' UTRs are rather short (3-10 nucleotides), the length of 5'UTRs in *S. cerevisiae* broadly varies. Comprehensive studies were carried out to determine the 5'UTR length of genes in the *S. cerevisiae* genome by rapid amplification of cDNA ends (5'RACE) and RNA-sequencing [229,230]. Typically, the length of yeast 5'UTRs spans a range of 0 to 500 bp with an average length of 83 bp [229]. Additionally, it was shown that the 5'UTR length is correlated with different cellular functions [229,231,232]. For example, genes associated with rRNA processing and protein folding seem to have shorter 5'UTRs than average, while genes playing a role in cell wall organization and metabolic processes tend to have long 5'UTRs, mainly due to the fact that these genes need considerable regulation. Next to the divergence in UTR length, the lack of a real consensus sequence to start translation in yeast, such as the Shine-Dalgarno sequence in *E. coli*, is a barrier in unraveling 5'UTR functionality. The 'ideal' context around the AUG for translation initiation is commonly depicted as the Kozak sequence [233,234], but still no unity hereabout exists, as other studies demonstrated that the nucleotide context around the start codon is far from conserved [235,236]. Furthermore, low correlations were found between the nucleotide distribution adjacent the start codon

and the gene translation rate [235]. Yeast 5'UTRs can also possess single or multiple regulatory elements such as secondary structures [237], upstream AUGs (uAUGs), upstream open reading frames (uORFs) [238] and internal ribosomal entry sites (IRES) to name a few [228] (Figure 4.1). The complexity of how these regulatory elements interact with each other, together with the diversity in 5'UTR length and the lack of an appropriate translational initiation consensus motif makes the engineering of 5'UTRs with a predictable outcome on protein expression a cumbersome quest.

Nevertheless, pioneering attempts hereto were already undertaken in *S. cerevisiae*. Dvir and coworkers made a large-scale library of the *RPL8A* 5'UTR, by randomizing positions -10 to -1 relative to A (pos. +1) of the AUG, of which 2041 variants were accurately sequenced and protein abundance was determined [23]. Next, predictive features of the 5'UTR were distilled using Least Absolute Shrinkage and Selection Operator (LASSO) regression on each training set determined via a 10-fold cross-validation scheme. Hereafter, robust features appearing in all 10-fold cross validations were extracted and subjected to best-subset regression analysis of which 13 predictive features were found, explaining 68% of the expression variation in the library. In addition, very recently a random library of half a million 50 bp 5'UTRs was constructed to train a convolutional neural network (CNN) model that had the power to predict the translational efficiency of 5'UTRs [239]. Obviously, these findings show the huge interest in predictive computational design methods to balance eukaryotic pathways at the translational level. However, the aforementioned studies only focused on 5'UTRs combined with one gene (coding sequence) and one promoter. At this moment, it is unknown if such a predictive model is generally applicable in contexts with other promoters and other coding sequences. Therefore, we developed the 'yUTR calculator' for *S. cerevisiae*, a data driven approach which can be used for the *de novo* design of 5'UTRs with a predictive outcome on translation initiation rates, applicable in combination with different promoters and different coding sequences. As a starting point, we used data of the 2041 5'UTR sequences of Dvir *et al.* [23] for the development of a partial least square (PLS) regression model. Next, via forward engineering, completely new 5'UTRs leading to protein abundances within a specified target range were created. Furthermore, to examine the universality of our yUTR calculator, it was validated in several representative case studies. As such, the developed yUTR calculator forms a basis toward custom, accurate predictive translational tuning in eukaryotic hosts.

4.3 MATERIAL AND METHODS

All products were purchased by Sigma Aldrich (Diegem, Belgium) unless otherwise stated.

4.3.1 Strains and media

S. cerevisiae BY4742 (*Matα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*, Euroscarf, University of Frankfurt, Germany, [205]) was used as yeast expression host. All yeast strains derived from this strain are listed in Supplementary Table S.2.1. Yeast cultures were grown in synthetic defined (SD) medium consisting of 0.67% YNB without amino acids, 2% glucose (Cargill, Sas van Gent, The Netherlands) and selective amino acid supplement mixture without uracil (CSM – URA, MP Biomedicals, Brussel, Belgium). To solidify media, 2% agar noble (Difco, Erembodegem, Belgium) was added. Synthetic Feed-In-Time (FIT) fed-batch medium was used to evaluate the p-coumaric acid production strains. FIT synthetic fed-batch medium M-Sc.syn-1000 was ordered from M2P labs (Baesweiler, Germany). Prior to use, an enzyme mix (final concentration of 0.5% v/v) and a vitamin mix (final concentration of 1% v/v) was added to the Sc.syn Base solution.

One Shot TOP10 Electrocomp™ *E. coli* (ThermoFisher Scientific, Aalst, Belgium) was used for cloning procedures and for maintaining template plasmids. *E. coli* strains were cultured in Lysogeny Broth (LB) consisting of 1% tryptone-peptone (Difco), 0.5% yeast extract (Difco), 1% sodium chloride (VWR, Leuven, Belgium) and additionally 100 µg/ml ampicillin. For solid LB growth medium, 1% agar was added.

4.3.2 Plasmid construction

To evaluate the different 5'UTR libraries, four pTemplate vectors with a reporter or a pathway transcription unit (TU) were developed. pTemplate1 consists of the *RPL8A* promoter with its native 5'UTR in front of yECitrine (pKT140, Euroscarf [206]), pTemplate2 holds the *TEF1* core promoter (*i.e. TEF1* promoter without UAS) and its native 5'UTR in front of yECitrine, pTemplate3 has the *RPL8A* promoter with its 5'UTR in front of mTFP1 (Genbank DQ676819 [240], gBlock, Integrated DNA Technologies (IDT, Leuven, Belgium)) and pTemplate4 exists of the *TEF1* promoter (*i.e. TEF1* core promoter with UAS) and its native 5'UTR in front of the *Rhodobacter capsulatus tal1* gene (*RcTal1* [241], gBlock, IDT). All four plasmids are yeast low copy expression vectors consisting of a *CEN/ARS4* origin of replication (ori) (Euroscarf) and a *URA3* auxotrophic marker (Euroscarf). Transcription of

yECitrine, mTFP1 and *RcTal1* is terminated by the *ADH1* terminator. All pTemplate vectors, except pTemplate4, also contain a *TEF2p-mCherry-PGK1t* TU to correct for cellular background variation with mCherry (iGEM part BBa_J06504, plate 3, 17C, 2013). The *RPL8A* promoter, *TEF2* promoter and *TEF1* (core) promoter, and the *ADH1* terminator and *PGK1* terminator were picked up from the BY4742 genome (Genbank JRIR00000000). A schematic overview of the pTemplate plasmids is given in Supplementary Figure S.2.1. For the construction of the different p_yC vectors, linear DNA was amplified using PrimeSTAR HS DNA polymerase (Takara, Westburg, Leusden, The Netherlands) using a 10 000x dilution of the pTemplate plasmids. After purification with the innuPREP PCRpure Kit (Analytik Jena AG, Jena, Germany) plasmids were assembled using two-pieces CPEC [11]. Primers and DNA oligonucleotides containing the 5'UTR library sequences were ordered from IDT. Sequencing of the plasmids was performed via EZ-seq (Macrogen Inc., Amsterdam, The Netherlands). Plasmids were isolated from E. coli cultures using the innuPREP Plasmid Mini Kit (Analytik Jena AG). An overview of all plasmids used in this study is given in Supplementary Table S.2.1. All 5'UTRs tested are listed in Supplementary Table S.2.2.

The p_yC vectors were transformed in yeast BY4742 via the lithium-acetate method following a Microtiter Plate Transformation protocol [242]. After transformation, s_yC strains were selected on SD CSM – URA plates and confirmed by yeast colony PCR using *Taq* DNA polymerase (NEB, Bioké, Leiden, The Netherlands).

4.3.3 In vivo fluorescence measurements

To evaluate the influence of the different 5'UTRs on protein abundance, fluorescence measurements were performed in 96-well microtiter plates (MTP). s_yC strains were plated on SD CSM – URA plates and incubated at 30°C for three days. Four colonies per strain were inoculated in 150 µl SD CSM – URA medium in a sterile polystyrene flat-bottomed 96 well plate (Greiner Bio-One, Vilvoorde, Belgium) covered with a Breathe-Easy® sealing membrane (Sigma Aldrich). These pre-culture MTPs were grown for 24h on a Compact Digital Microplate Shaker (ThermoFisher Scientific) at 800 rpm and 30°C. Subsequently, these pre-cultures were diluted 150 times in 150 µl fresh SD CSM – URA medium and grown in sterile polystyrene black µclear flat-bottomed 96 well plates (Greiner Bio-One). These MTPs were grown for 30h (till stationary phase) on a Compact Digital Microplate Shaker (ThermoFisher Scientific) at 800 rpm and 30°C. Finally, fluorescence (FP) of the *S. cerevisiae* strains was measured using a SpectraMax M2/M2e microplate reader (Molecular Devices,

UK). For measuring mTFP1, yECitrine and mCherry fluorescence, excitation wavelengths of respectively 460 nm, 500 nm and 575 nm and emission wavelengths of respectively 500 nm, 540 nm and 620 nm were used.

For every strain, the yECitrine-to-mCherry ratio or the mTFP1-to-mCherry ratio was used as a measure for protein abundance (PA). The yECitrine-to-mCherry (PA_C) fluorescence ratio was calculated as follows:

$$PA_C = \frac{FP_{yECitrine}}{FP_{mCherry}} \tag{4.1}$$

The mTFP1-to-mCherry (PA_M) fluorescence ratio was calculated as follows:

$$PA_M = \frac{FP_{mTFP1}}{FP_{mCherry}} \tag{4.2}$$

Protein abundances were normalized by dividing every calculated protein abundance by the mean protein abundance of all strains for a given experiment. This was calculated as follows:

$$normalized\ PA_C = \frac{PA_{C,single\ strain}}{mean\ PA_{C,all\ strains}} \tag{4.3}$$

$$normalized\ PA_M = \frac{PA_{M,single\ strain}}{mean\ PA_{M,all\ strains}} \tag{4.4}$$

As such, in all plots the normalized protein abundances of the predicted and *in vivo* measured values are shown.

4.3.4 Model feature quantification

In total 13 features (variables), which were chosen based on literature [23], were calculated for every 5'UTR in the database of Dvir *et al.* (see Supplementary Table S.2.3 for detailed definitions). These 5'UTR features were categorized into four groups conforming with the classification of Dvir *et al.* [23]. Six features determined the nucleotide preferences at positions

-3 to -1 upstream of the AUG (AUG context) [a], five features checked for short k-mer sequences in the 5'UTR (short k-mer sequences), one feature searched for out-of-frame uAUG's (oof_uAUG's) and a last feature calculated the ensemble free energy (EFE) of the mRNA secondary structure (RSS). Except for the RSS and the number of out-of-frame uAUG's, all features were given a value of 1 if the concerned sequence motif was present or a value of -1 if not. For the uAUG feature, the effective number of oof_uAUG's present in the 5'UTR was given and for the RSS feature, the value of the EFE of the mRNA secondary structure of the 5'UTR and the first 50 base pairs of the coding sequence was calculated. Therefore, the intramolecular interactions in the RNA molecules were predicted using RNAfold [243]. This RNA secondary structure prediction algorithm is available in the Vienna RNA package [244] and was used with the default settings and the options –noLP –d2 and an accuracy of 10^{-100}. All feature calculations were performed via Python scripting on a Dell Latitude E6540 laptop.

4.3.5 Partial least squares (PLS) regression

The partial least square (PLS) regression was performed in R using the pls package [245]. All protein abundance data of the 2041 5'UTR sequences [23] were rescaled by dividing their values by the minimum protein abundance. In addition, the 5'UTR dataset was randomly split in a training set and a test set by a 5:1 ratio. The training set was used to create the PLS regression model; the independent test set was used to check the prediction capability of the model. The latent variables of the PLS model were determined using a 10-fold cross-validation (CV). Model validation was done using an independent test set to calculate the coefficient of determination (R^2). Prior to regression, predictor variables were scaled by dividing each variable by its standard deviation. The linear relationship between the response variable (*i.e.* protein abundance) and the 13 predictors is given in equation 4.5. Herein is **Y** the matrix of protein abundances, **X** the matrix of predictors (*i.e.* 13 model features), **B** the matrix with regression coefficients and **ε** an error matrix. In PLS regression, the matrix of predictors **X** is decomposed into an orthogonal score matrix **T** and loadings matrix **P**, which circumvents possible collinearities. Next, **Y** (*i.e.* protein abundance) is regressed on score matrix **T** and not on **X** itself. More specifically, the kernel PLS algorithm was used, described by Dayal *et al.* [246].

[a] The A of the AUG is seen as position +1. All nucleotides preceding the start codon are numbered relative to this position ending with position -1 for the nucleotide in front of the start codon.

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{B} + \boldsymbol{\varepsilon} \tag{4.5}$$

$$\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{P} \tag{4.6}$$

4.3.6 Search algorithm for de novo design of 5'UTRs

An iterative search algorithm was developed for the creation of novel 5'UTRs. It starts with the *ad random* creation of 100 degenerated 5'UTRs between positions -10 to -1. All 13 5'UTR features of these sequences are analyzed and protein abundance is predicted with the PLS model. The scatter of protein abundances of the 5'UTR candidates is evaluated by examining how well these 5'UTRs cover the wanted protein abundance range. Therefore, the range of protein abundances is divided in equal bins and it is checked how proper the predicted 5'UTR protein abundance is fitting each bin. In an optimal situation, the protein abundance of a candidate should be in the middle of a bin. Next, during the first iteration, these 100 candidates are subjected to degenerations, mutations and recombinations. For all these generated 5'UTR candidates, features are determined, protein abundance is predicted and scatter is evaluated. Subsequently, the 100 best candidates are pooled and a second iteration is started. In conclusion, after several iterations, libraries with 5'UTR candidates that cover the entire desired protein abundance range are chosen and reported (*i.e.* the best 5'UTR candidate per bin is reported). The desired range of protein abundances, the number of bins and the number of iterations is specified beforehand.

4.3.7 Cultivation of p-coumaric acid production strains

For the growth experiments with the p-coumaric acid strains, three biological replicates per strain were inoculated from agar plate in 200 µl selective SD medium in a sterile µclear, flat-bottomed, white 96-well microtiter plate (Greiner Bio-One, Vilvoorde, Belgium) enclosed by a Breathe-Easy® sealing membrane (Sigma-Aldrich). These pre-culture MTPs were grown for 24h on a Compact Digital Microplate Shaker (ThermoFisher Scientific) at 800 rpm and 30°C. For the main cultivation experiments MTPs with air-penetrable sandwich cover (EnzyScreen, Heemstede, The Netherlands) were used. 50 µl of the pre-culture was used for inoculating 500 µl medium in 96 deep-well MTPs (EnzyScreen). All cultivations were carried out for 72h at 30°C and 350 rpm (2.5 cm orbit). At the end of cultivation, the optical density was measured at 600 nm (OD600) by diluting 15 µl culture in 135 µl deionized water in a µclear, flat-bottomed, black 96-well microtiter plate (Greiner Bio-One). The

OD600 was determined in a TECAN Infinite® 200 PRO (Tecan) MTP reader. Afterwards, cultures were spun down and the supernatant was used for metabolite detection and quantification using Ultra Performance Liquid Chromatography (UPLC).

4.3.8 Detection and quantification of p-coumaric acid

p-coumaric acid was measured using a Waters Acquity UPLC connected to a UV detector and equipped with a Kinetex® 2.6 µm Polar C18 column (Phenomenex, Utrecht, The Netherlands) operated at 30°C. A gradient method with two eluents, *i.e.* 13 mM trifluoroacetic acid (TFA) (A) and pure acetonitrile (ACN) (B), with a flow rate of 0.6 ml/min was used. The UPLC method started with 10% of eluent B, followed by a linear increase to 23% of eluent B (0 – 2.5 min) where its fraction was subsequently further increased to 70% (2.5 – 5.0 min). Next, the fraction was maintained at 70% of eluent B (5.0 – 6.0 min), finally the fraction of eluent B was decreased from 70% to 10% (6.0 – 8.0 min). p-coumaric acid was detected at 290 nm with a retention time of 2.3 min. The peak area was integrated with OpenChrom® and concentrations were determined from a p-coumaric acid standard curve. This standard was HPLC grade (> 95% purity) and purchased from Sigma-Aldrich.

4.3.9 Data analysis

All calculations were done in Python using the Python Data Analysis Library (Pandas). Unless mentioned otherwise, error bars represent the standard error of the mean (n = 4). All coefficients of determination were calculated using the hydroGOF package in R or the statsmodels package in Python. The scipy.stats package in Python was used to determine p-values via a two-sided T-test. In all cases, a significance level of 0.05 was applied.

The 95% confidence interval of the linear regression was used to explain the accuracy between the predicted protein abundances and the *in vivo* measured protein abundances. Designed 5'UTR sequences laying in the 95% confidence interval were seen as accurately fitting the desired target level predicted by the model. As such, the percentage of 5'UTRs within the 95% confidence interval gives an idea of the accuracy between predicted and measured protein abundance.

4.4 RESULTS AND DISCUSSION

The suitability of regression methods in engineering biological systems has already been proven in earlier studies [247–249]. Furthermore, in *S. cerevisiae*, it was shown that such models can be useful to understand and predict the influence of sequence patterns on gene expression [195,196,199,215,250,251]. Since a lot of these studies focused on the transcriptional landscape, a computational approach was developed here to create *de novo* 5'UTRs which have a predictive influence on protein abundance in yeast.

4.4.1 Development of the yUTR calculator

In the available data set of 5'UTR sequences, 2041 5'UTR sequence variants with their respective protein abundances are presented [23]. Before starting with PLS regression, 13 defined features of every 5'UTR candidate (Supplementary Table S.2.3) were calculated. As such, an output file was generated containing for each 5'UTR the nucleotide sequence, the measured protein abundance and all 13 calculated 5'UTR features. This output file is available on GitHub (https://github.com/DeMeylab/2018---yUTR-calculator).

Next, this data set was randomly divided in two subsets. One subset was used as a training set to build the model and contained data of 1633 5'UTRs, the other subset was an independent test set to validate the model and included data of 408 5'UTRs (Figure 4.2A). The model was calibrated using the training set and incorporated all 13 features describing protein abundance. By using 10-fold cross-validation, 4 latent variables were retained based on the root mean squared error of prediction and the percentage of explained Y variance (Supplementary Figure S.2.3). With 4 latent variables, the PLS model covers 36.65% of the X variance, which explains 67.34% variance of the response variable Y. The coefficient of determination (R^2), a measure for the model efficiency, was 0.67 (Supplementary Figure S.2.4).

To evaluate the quality of this PLS model, the independent test set of 408 5'UTRs was used. Although a small fraction of 5'UTRs had a much higher protein abundance as predicted by the model, specifically in the predicted PA region lower than 2, an R^2 of 0.73 was obtained for this validation set, indicating that the PLS model has the potential to successfully predict protein abundance (Figure 4.2B). In comparison, a recently developed CNN model described by Cuperus *et al.* and using a data set of 50 bp 5'UTRs led to an R^2 of 0.62 [239]. In Supplementary Figure S.2.5, the estimated regression coefficients of the 13 5'UTR features

are represented, illustrating the most influential factors in the PLS model. In addition, the biplot of the first two components (*i.e.* latent variables) and the cumulative loadings of the four components are given in Supplementary Figure S.2.2 and Figure S.2.6, respectively. Based on the calculated regression coefficients, it is clear that the two predictors AG_in_min3 and oof_uAUG have a huge impact on protein abundance. The regression coefficients of AG_in_min3 and oof_uAUG in our model are respectively positive and negative, indicating that the presence of a purine at position -3 and the absence of uAUGs lead to high protein levels. Similar effects on protein expression in yeast have already been described [228,252,253]. Other observations are the positive influence on protein abundance of an adenine at position -1 and the presence of an adenine dimer at position [-3,-2], also found in the combined model of Dvir and coworkers [23]. On the other hand, the presence of a CACC 4-mer in the 5'UTR negatively affects protein expression. Beside the sequence features like AUG context and short 4-mer subsequences, the mRNA secondary structure in particular, calculated by the ensemble free energy (dG_EFE) of the mRNA, is a prevailing feature in the model (Supplementary Figure S.2.5). Other attempts showed the regulatory role of secondary structures within the 5'UTR [237,254–256], rather than focusing on the secondary structure of the entire mRNA. To this end, in an earlier approach by Crook *et al.* [257], the process of translation initiation was modeled by assuming that the initiation complex is a particle that has to surmount different secondary structures within the 5'UTR, each having their own free energy of folding. In this respect, 5'UTR hairpins were also recently used to predictably tune protein expression in yeast [209]. As such, especially for longer 5'UTRs and in addition to focus on the whole mRNA secondary structure alone, modifying the position and the strength of secondary structures in the 5'UTR preceding the start codon, could be interesting extra parameters (*i.e.* beside the AUG context, subsequences and uAUGs) for our model in the future to improve the prediction and rational design of novel 5'UTRs with user defined functions.

The three most influential 5'UTR features in our model (*i.e*. adenine at position -3, effect of secondary structure and uAUGs) were also observed to be important factors when analyzing the large 5'UTR data set of Cuperus and coworkers [239]. Moreover, the effect of the 5-mer ranging from position -5 to -1 on protein expression, indicated as the Kozak sequence, was additionally assessed in their model and several Kozak sequences were linked with strong 5'UTRs. Beside the position dependent features and secondary structure of the 5'UTR captured in our linear model, Cuperus *et al.* used a CNN to predict protein expression

from random 5'UTR sequences which has the additional advantage to cope with nonlinear interactions between features. As such, for example, the influence of uORFs on protein expression could be more accurately learned compared to a linear model. To this end, Cuperus and coworkers observed that uAUGs in frame with the real start codon and without a stop codon in front of this AUG caused a minor reduction in gene expression compared to out of frame uAUGs [239]. However, since computational approaches based on CNN are complex, a lot of data is needed to achieve a highly predictive model. Indeed, it was observed that the predictive power of the CNN model decreased with smaller training sets and that all 50 nucleotides of the 5'UTR had to be included as with only the 10 nucleotides preceding the start codon, bad predictions were determined ($R^2 = 0.097$) [239]. To this end, with the small data set used here which is based on only 10 nucleotides adjacent to the AUG, it can be concluded that PLS regression is a useful method in our study ($R^2 = 0.73$) to find a relationship between protein abundance and the 13 sequence features in the 5'UTR.

Hence, this model was used to develop a yUTR calculator that can design *de novo* 5'UTR sequences for *S. cerevisiae* with a predicted outcome on protein abundance (Figure 4.2C). The search algorithm described in Material and Methods is the core of the yUTR calculator. Its parameters were specified as follows: the desired protein abundance was set between 2 and 8 and the number of bins was fixed at 8. As the model had difficulties to accurately predict protein abundances lower than 2 (Figure 4.2B), it was decided to pin the lowest set point at 2 instead of 1. By choosing 8 bins, eight 5'UTR candidates are sufficient to completely cover the desired protein abundance range.

Using the yUTR calculator, three different 5'UTR libraries (*i.e.* UTRa, UTRb and UTRc) of 16 candidates were designed for various contexts of promoters, 5'UTRs and coding sequences. To examine if scatter accuracy was improved when applying more repetitions of the search algorithm, the yUTR calculator was ran 3 times with a setting of 250 iterations and 2 times with 500 iterations for the generation of every 5'UTR library. In fact, no difference in library scatter accuracy was observed. From these results, two libraries of eight 5'UTR candidates with visually the most equal spreading (Supplementary Figure S.2.7) were selected to form libraries UTRa, UTRb and UTRc. An overview of the sequences in the 5'UTR libraries and the corresponding expression plasmids is available in Supplementary Table S.2.2 and Table S.2.1. Due to cloning issues expression vectors p_yC^{III}-4, p_yC^{III}-16 and p_yC^{IV}-6 could not be constructed and so were not taken into account in the *in vivo* library evaluation. For the

latter, the yECitrine-to-mCherry (PA_C) or mTFP1-to-mCherry (PA_M) ratio was determined and used as measure for 5'UTR strength. The native 5'UTRs of the *RPL8A* promoter and the *TEF1* promoter served as reference for the *in vivo* analysis of the novel 5'UTR libraries (sTemplate1-3, Supplementary Table S.2.1). Their predicted protein abundance values were determined via reverse engineering with the constructed PLS model.

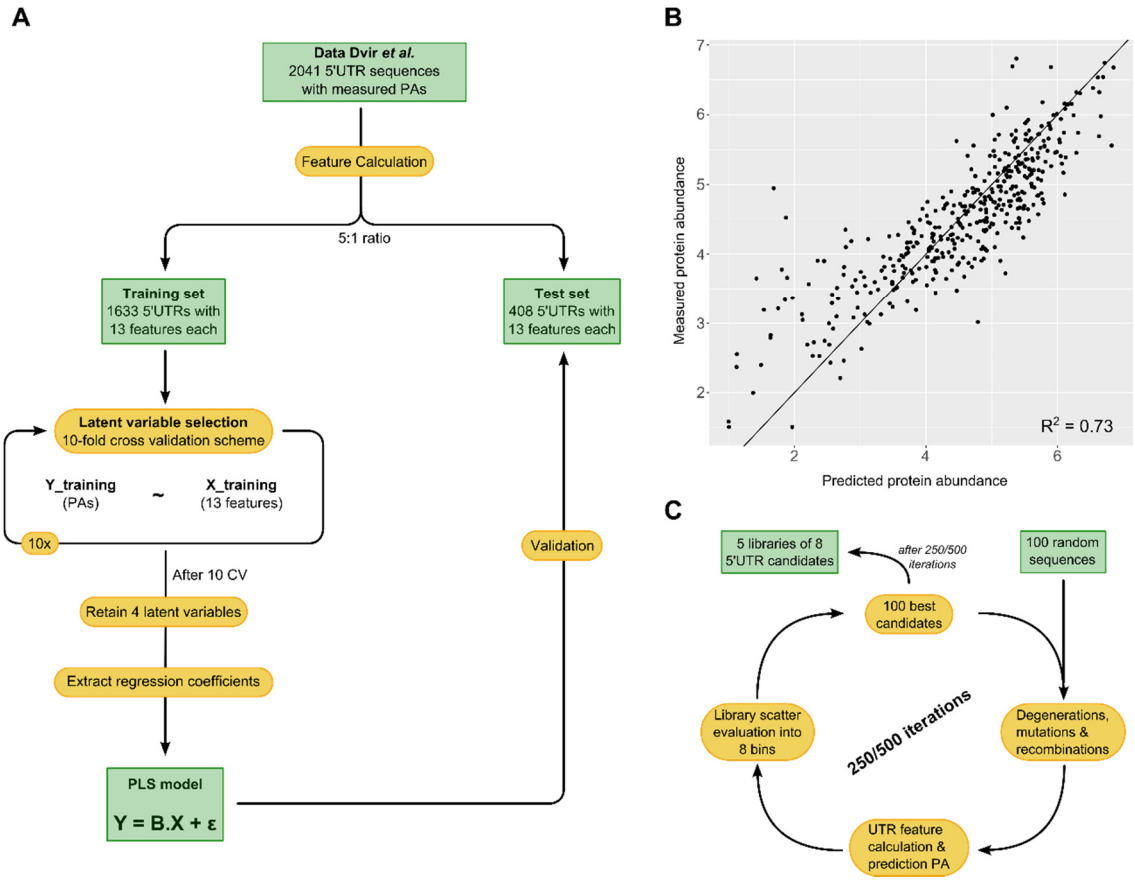

Figure 4.2: (A) Flow diagram of the construction of our PLS model. After determination of the feature values for each 5'UTR, the data set is split in a training and test set. The training set is used to learn a PLS regression model by using a 10-fold cross validation (CV) scheme, after which 4 latent variables were retained (pls package in R [245]). Finally, the PLS model with 4 components was further validated on an independent test set to examine its potential to predict protein abundances (PAs). (B) Validation of the PLS regression model. The model uses 13 features of the 5'UTR of *Saccharomyces cerevisiae* (Supplementary Table S.2.3) to predict protein abundance. This plot represents the measured [23] versus the predicted protein abundance, calculated via the PLS model, for the test set of 408 5'UTRs. As measure for the model efficiency, a coefficient of determination (R^2) of 0.73 was obtained. (C) Schematic overview of the iteration process used in the 'yUTR calculator' to design *de novo* 5'UTR sequences with predicted protein abundances that fully cover a specified protein abundance range.

4.4.2 The yUTR calculator compared to the Dvir model

As described above, using the yUTR calculator, three different 5'UTR libraries, UTRa, UTRb and UTRc, were designed for various contexts of promoters, 5'UTRs and coding sequences. Library UTRa was created by forward engineering using the *RPL8A* promoter, its 17 bp 5'UTR and the yECitrine coding sequence (CDS) (Figure 4.3A, Supplementary Figure S.2.8). This context is equal to the design of Dvir and coworkers [23] and was developed to verify if our yUTR calculator, used for the creation of *de novo* predictive 5'UTRs, had a similar predictive power as their model. None of the 16 novel 5'UTRs of library UTRa did appear in the original data set of 2041 5'UTR sequences used to build and train our PLS model. *In vivo* evaluation of the s_yC^{I} strains and reference strain sTemplate1, showed an accuracy of 53% and an R^2 of 0.70 was obtained (Figure 4.3A, Supplementary Figure S.2.14). As expected, this is in line with the results of Dvir and coworkers (*i.e.* R^2 = 0.69) [23], since the same 13 features were taken into account in our PLS model. In addition, the library resulted in a 2.0-fold differentiated expression landscape, in line with the predicted 2.5-fold range. Six strains had a significantly higher (p-values < 10^{-3}), and five strains a significantly lower (p-values < 0.05) yECitrine activity than the reference strain containing the native *RPL8A* 5'UTR.

When looking in detail to some sequence motifs in library UTRa, some earlier described patterns reappear. An out-of-frame (-1) uAUG was found in UTRa2, 4, 6 & 8, causing weak yECitrine activity (Supplementary Figure S.2.14). Indeed, the presence of uAUGs, a feature captured in the PLS model and again confirming the latter's predictive power as low protein abundances were predicted, causes weak protein expression because part of the ribosomes already start translation at these uAUGs[228,239,253,258]. The CAAG 4-mer, a motif not taken into account in our model, found in UTRa11 and UTRa15 was previously related to increased *HIS3* protein expression [239]. Here, UTRa11 led to a strong output, while for UTRa15 this was not the case (normalized PA of 1.32 ± 0.01 versus 0.82 ± 0.02, respectively). However, in contrast to UTRa11, UTRa15 has a T at position -3, confirming the very strong negative effect on protein expression of this nucleotide on position -3 in the yeast 5'UTR [235,259,260].

Altogether, these findings prove that our yUTR calculator can be used to develop *de novo* 5'UTR sequences with a foreseen effect on protein expression. Also, 5'UTRs with higher translation initiation rates than the native 5'UTR in the same context were obtained.

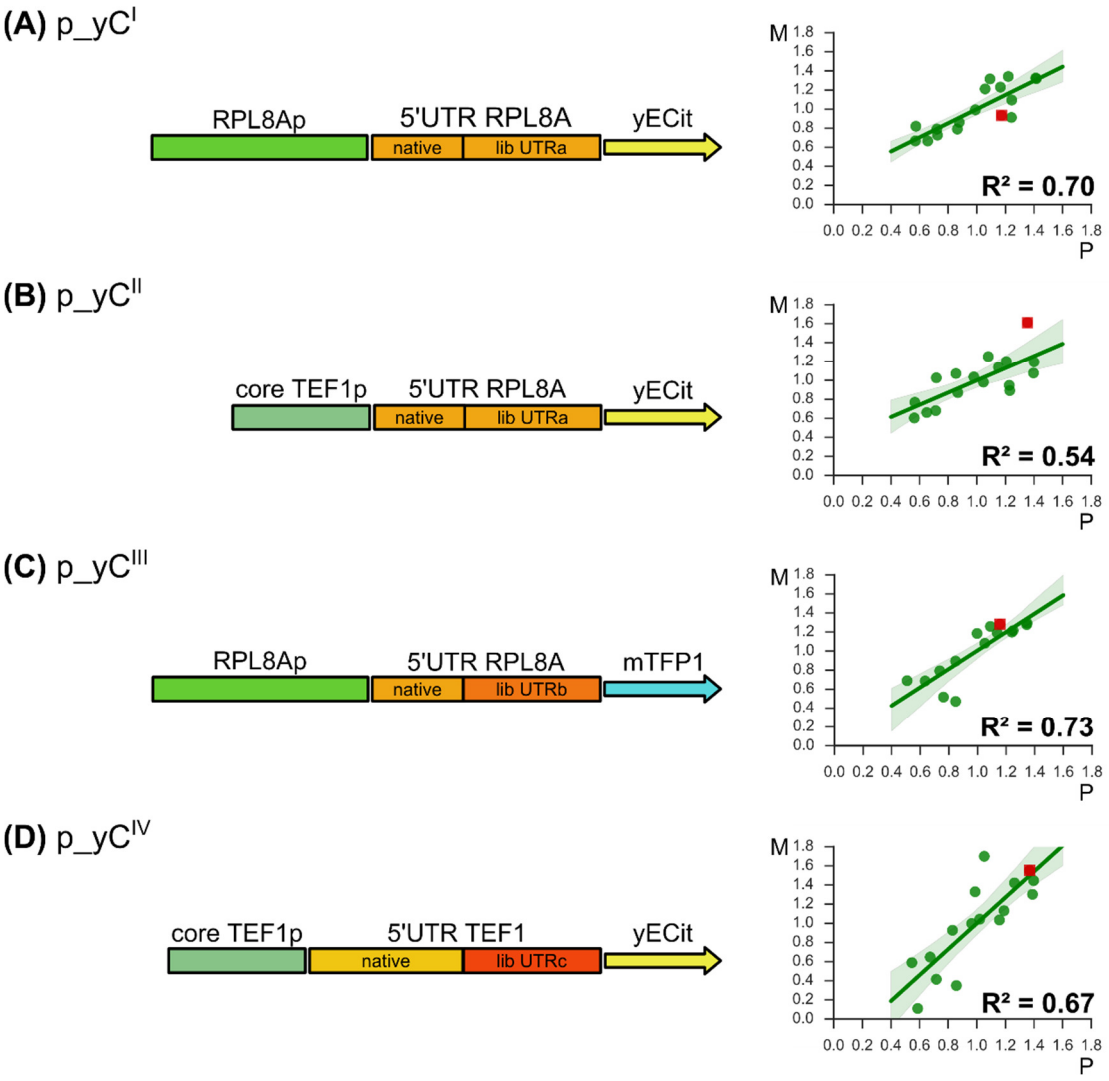

Figure 4.3: *In vivo* analysis of *de novo* designed 5'UTR sequences for four different experimental set-ups. Shown left is a schematic representation of the transcription unit (TU) used to evaluate a particular calculated 5'UTR library. On the right, OLS regression plots are shown between normalized predicted PA calculated by our 'yUTR calculator' (P) and normalized measured PA determined *in vivo* (M). Dots show the mean PA of four biological replicates for the 16 strains and their reference strain (to be complete only 14 strains and 15 strains were evaluated for p_yC^{III} and p_yC^{IV}, respectively). The linear fits are represented with their respective coefficient of determination (R^2) and their 95% confidence intervals. In every regression plot, the red square dot represents a reference strain containing the promoter with its native 5'UTR. (A) TU existing of the native yeast *RPL8A* promoter, a 5'UTR existing of the first 7 original nucleotides of the *RPLA8* 5'UTR with the designed 5'UTR library UTRa and the yECitrine CDS. As a reference strain, sTemplate1 was used. (B) TU existing of the native yeast core promoter of the *TEF1* gene, a 5'UTR consisting of the first 7 original nucleotides of the *RPLA8* 5'UTR with the 5'UTR library UTRa and the yECitrine CDS. Strain sTemplate2 was used as reference here. (C) TU is the native yeast *RPL8A* promoter, a 5'UTR existing of the first 7 original nucleotides of the *RPLA8* 5'UTR with the calculated 5'UTR library UTRb and the mTFP1 CDS. The reference strain here was sTemplate3. (D) TU consisting of the native yeast core promoter of the *TEF1* gene, a 5'UTR consisting of the first 23 original nucleotides of the *TEF1* 5'UTR with the novel 5'UTR library UTRc and the yECitrine CDS. Strain sTemplate2 was the reference here.

4.4.3 Universal applicability of the yUTR calculator

In a next step, library UTRa was used in a different transcriptional context, with the *TEF1* core promoter instead of the *RPL8A* promoter (Figure 4.3B, Supplementary Figure S.2.9), to verify if the promoter has an influence on the predictive outcome. The core promoter was chosen instead of the native *TEF1* promoter since it was reported that the core promoter largely contributes to the activity of the entire promoter [196]. To improve protein expression, the core promoter could be extended with different upstream activated sequences (UASs) in future research, as it was shown that UASs are transcriptional amplifiers [110]. Evaluating strains s_yC^{II} an R^2 of 0.54 was derived by OLS regression (Figure 4.3B, Supplementary Figure S.2.15) indicating that the predicted protein abundance values are less in line with their *in vivo* behavior compared to the context for which they were designed (*i.e.* for combination with the *RPL8A* promoter). 10 out of 17 (59%) of the designed sequences had measured protein levels within the 95% confidence interval of the linear fit and the order of *in vivo* measured protein abundance was mostly unchanged for the higher and lower strength 5'UTRs (*e.g.* p = 6.44E-6 and p = 6.60E-4 for a pairwise comparison between UTRa3 and UTRa6 in strains s_yC^{I} and s_yC^{II}, respectively (Supplementary Figure S.2.15)). For the intermediate strength 5'UTRs no definite conclusion hereabout could be taken due to the large standard errors (Supplementary Figure S.2.15). In addition, just like for the s_yC^{I} strains, a 2.1-fold variation in abundance was observed. Though generally these observations could imply that another promoter does not influence the 5'UTR and can be used as a discrete DNA building block to enlarge or reduce mRNA abundance [261], the mean yECitrine-to-mCherry ratios (Eq. 4.1) did not differ significantly (p = 0.6307) for both s_yC^{I} and s_yC^{II} populations and a lower measure for the model efficiency (R^2) was obtained. This suggests that transcription and translation are inseparable and care must be taken to see the promoter and 5'UTR as discrete DNA building blocks [253,262].

Library UTRb was designed with the *RPL8A* promoter, its 17 bp 5'UTR and the mTFP1 CDS, a fluorescent reporter with a totally different CDS of the one of the yECitrine reporter (Figure 4.3C, Supplementary Figure S.2.10). With this composition, we were able to validate if our model can predict protein abundances when another coding sequence is given. Since the effect of the coding sequence is captured in the dG_EFE feature of our model, it was expected that the yUTR calculator could generate new 5'UTRs for library UTRb with acceptable predictability. Indeed, an R^2 equal to 0.73 was obtained, which confirms on the

one hand that our model can accurately predict differences in protein expression with another coding sequence. In addition, the measured protein abundances covered a 2.8-fold expression range (Figure 4.3C, Supplementary Figure S.2.16) and an accuracy of 67% between the predicted and measured protein levels was obtained. In contrast to expression library p_yC[I], no 5'UTRs led to a significantly stronger mTFP1 expression than the reference strain sTemplate3 (p-values > 0.05). Again, the strong and weak behavior of some 5'UTRs could be explained by earlier described regulating sequence patterns [239]. For example, despite the presence of a C at position -3, UTRb3 led to one of the highest mTFP1 outputs (normalized PA of 1.26 ± 0.03) through the presence of the AAGA, ACAA and TACA 4-mers, which suggests that when multiple strong motifs are combined, this can overcome the negative effect of the pyrimidine at position -3. Additionally, out-of-frame uORFs with a stop codon before the primary ORF were determined in UTRb2 and UTRb6, having a low PA of 0.47 ± 0.01 and 0.52 ± 0.02, respectively. Together with out-of-frame uAUGs in UTRb8, 10, 12 and 14, leading to reduced protein abundances of, for example, 0.69 ± 0.03 and 0.79 ± 0.06 for respectively UTRb8 and UTRb12, this again illustrates the unfavorable impact of uAUGs/uORFs [258,263–265].

Finally, it was tested if our computational approach could be generalized toward other promoters and longer 5'UTRs. Therefore, library UTRc was created by forward engineering with the *TEF1* core promoter, its 33 bp 5'UTR and the yECitrine reporter (Figure 4.3D, Supplementary Figure S.2.11). As the model was originally developed based on protein abundance data from 5'UTRs in the *RPL8A* promoter and particularly its short 17 bp 5'UTR context, the *de novo* design of library UTRc was the most ambiguous calculation, since we used the model now out of the scope for which it was calibrated. The results of library UTRc showed a remarkable good fitness between the predicted values and the measured fluorescence output (R^2 equal to 0.67, Figure 4.3D, Supplementary Figure S.2.17) and for 63% of the generated 5'UTR candidates, measured protein levels were within the 95% confidence interval, indicating that the model can cope with the longer 5'UTR of the *TEF1* promoter. Compared to the reference pTemplate2, only 5'UTR sequence UTRc5 caused a significantly higher protein expression (p = 0.026). This is in line with earlier results, as it turned out that UTRc5 had a TATA 4-mer together with the ATAAG Kozak sequence, one of the strongest Kozak sequences reported [239]. Again, as previously reported and proven several times in our study, the nucleotide at position -3 plays a very decisive role in controlling mRNA translation. Whereas UTRc1, 2, 10, 12, 14 & 16 all had the enforcing 4-

mer TATA [239], only UTRc12 and UTRc16 had the negatively influencing T at position -3, clearly leading to lower protein abundances compared to their A containing counterparts (*e.g.* PA for UTRc16 was 0.35 ± 0.03 which significantly differs from UTRc1, 2, 10, 14, p-values < 10^{-4}, Supplementary Figure S.2.17). It also has to be noted here that the 23 nucleotides at the 5' end of the 5'UTR were kept constant. If these had a big effect on translation, little variation in protein expression would be noticed. Yet, a 15-fold difference in protein levels was measured (Figure 4.3D, Supplementary Figure S.2.17). Since only the 10 nucleotides in front of the start codon were modified, this implies that these 10 nucleotides before the AUG have a big effect on translation and as such are sufficient to reliably predict yeast protein expression. Indeed, it has been demonstrated that translation starts more efficiently when the start codon is surrounded by a specific context [233,235,266]. This hypothesis is also supported by the study of Cuperus *et al.* where out-of-frame uAUGs or uORFs in the 5'UTR led to the strongest reduction in protein expression when they were present near the start codon [239].

4.4.4 Protein coding sequence influence and reverse engineering

It was demonstrated in prokaryotes that reusing the same RBS sequence in front of another coding sequence does not work reliably [30,180], and that the first 50 to 100 nucleotides of the protein coding sequence strongly influence mRNA secondary structure in eukaryotes [23,239]. This effect of mRNA secondary structures was captured in the dG_EFE feature of our model, and was proven to work well (R^2 = 0.73 for s_yC[III]). To evaluate however the impact of characterized 5'UTRs on a different coding sequence, the fluorescence output was analyzed when library UTRa, which was designed in the context of the yECitrine CDS, was placed in front of the mTFP1 CDS (Figure 4.4B, Supplementary Figure S.2.12).

The aforementioned results already showed that our model is able to predict variation in protein expression (R^2 = 0.70, Figure 4.4A top), at least when the effective coding sequence is used for the predictions. However, when library UTRa was used in front of another protein CDS like for example mTFP1, the predicted values did not match very well with the measured output (R^2 = 0.35, Figure 4.4A bottom). As such, reusing a 5'UTR developed for one specific protein in front of another protein will probably not work reliably in yeast. This definitely demonstrates the significance of specifying the right coding sequence in the design of novel 5'UTRs.

Finally, to illustrate the reverse engineering capability of the PLS model, predicted protein abundances were recalculated for every 5'UTR in library UTRa when preceding the mTFP1 reporter (Figure 4.4B, Supplementary Table S.2.1). The coefficient of determination (R^2) for the whole library then was 0.69, suggesting a good predictive capacity of the model for existing 5'UTRs, provided their strength is recalculated with the effective CDS (Figure 4.4C, Supplementary Figure S.2.18). Indeed, also for every reference strain, a good prediction of the expression output was obtained by reverse engineering.

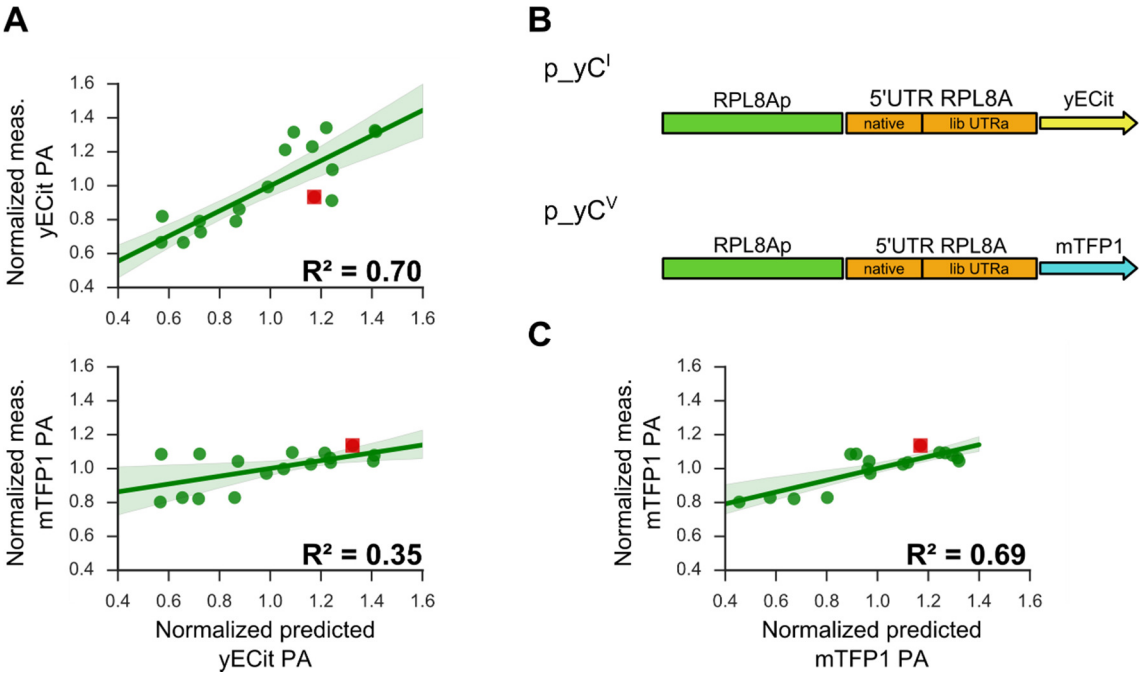

Figure 4.4: Influence of the protein coding sequence on the expression levels of different proteins. (A) Regression plots from 5'UTR library UTRa preceding the yECitrine (p_yC^I, top) and mTFP1 (p_C^V, bottom) CDS, respectively. The square red dots represent the reference strains sTemplate1 (top) and sTemplate3 (bottom). (B) Schematic overview of the TU existing of the native yeast *RPL8A* promoter, a 5'UTR existing of the first 7 original nucleotides of the *RPLA8* 5'UTR with the 5'UTR library UTRa and the yECitrine (p_yC^I, top) or mTFP1 (p_yC^V, bottom) CDS. (C) Regression plot from 5'UTR library UTRa preceding the mTFP1 CDS after recalculating the predicted protein abundance by reverse engineering. For all regression plots, the 95% confidence interval is given.

4.4.5 Proof of concept: reliable p-coumaric acid production

To evaluate the adaptability of the yUTR calculator beyond the use of fluorescent reporters, *de novo* 5'UTRs were developed for the *Rhodobacter capsulatus tal1* (*RcTal1*) coding sequence and tested for their predictable effect on p-coumaric acid production. The bacterial RcTal1p is responsible for the conversion of tyrosine into p-coumaric acid, which is an important precursor molecule for a lot of secondary metabolites such as stilbenoïds

(*e.g.* resveratrol) [5,267] and flavonoids (*e.g.* naringenin) [65,212]. With their promising bioactive properties, these compounds attain huge attention for usage in the pharmaceutical and food industry making their secured and defined supply essential. To this end, sustainable production with micro-organisms is a valuable alternative for the current extraction – and chemistry based production processes. Nevertheless, fine-tuning all steps in a heterologous pathway and the native metabolism is still needed to obtain an economic feasible microbial production process.

In this respect, a TU to functionally express the *RcTal1* CDS existing of the native *TEF1* promoter (*i.e.* core promoter including its upstream activating sequence) and the *ADH1* terminator was designed (Figure 4.5, Supplementary Figure S.2.13).

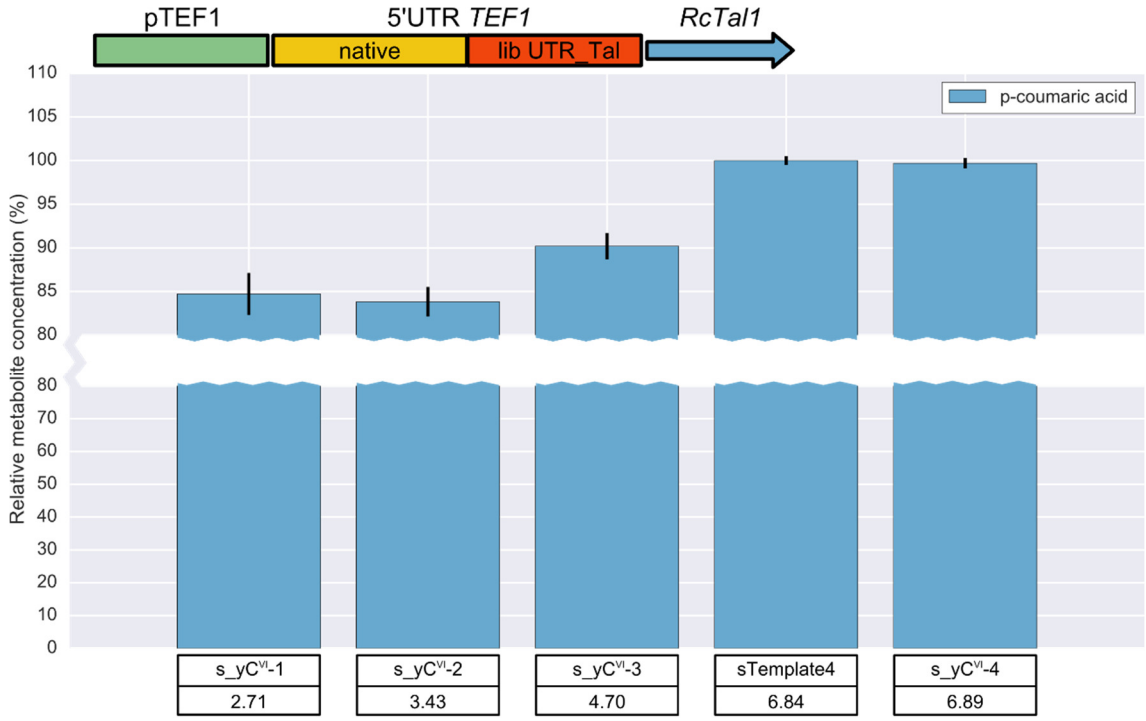

Figure 4.5: Titers of p-coumaric acid after introduction of 5'UTRs with a predicted outcome on protein abundance. Production titers were normalized against the highest producing strain sTemplate4 with the native TEF1 5'UTR. The strain names with corresponding predicted protein abundances are indicated in the black boxes (Supplementary Table S.1). Strains were grown for 72h in Feed-In-Time fed-batch medium. Error bars represent the standard error of three biological replicates.

To reliably alter the *RcTal1* translation initiation, our yUTR calculator was ran two times (250 iterations) for the design of novel 5'UTRs in front of the *RcTal1* CDS. Four out of the

sixteen rendered 5'UTR sequences were selected based on their predicted protein abundance for the construction of TUs; two weak, one intermediate and one strong 5'UTR. As a reference, the native *TEF1* 5'UTR was also taken into account and protein abundance was predicted through reverse engineering.

Production experiments in deep-well MTPs on FIT medium led to altered p-coumaric acid titers. The obtained results were very promising since 5'UTRs with weak and 5'UTRs with high predicted protein abundances effectively led to lower and higher production titers of p-coumaric acid, respectively (Figure 4.5). Also the native *TEF1* 5'UTR with a high predicted protein abundance calculated via reverse engineering caused high *in vivo* generated p-coumaric acid amounts which again confirmed the accurate reverse engineering capacity of the model. Altogether, these results prove the applicability of the yUTR calculator to reliably engineer expression levels of pathway genes that lead to metabolite titers generally in line with their predicted protein abundance. To this end, the yUTR calculator contributes to the further development of reliable forward engineering approaches for pathway balancing in *S. cerevisiae*, as such reducing the development times of profitable yeast cell factories.

4.5 CONCLUSION

With the ever increasing complexity of biological designs, the need for tools to reliably control (heterologous) pathway behavior and even understand the underlying working mechanisms has become inevitable. Modifying the translation initiation rates of genes has been proven to be a valuable tool in prokaryotes [30,180,181] and recently also in eukaryotes [23,239]. In this study, we developed a PLS model based on earlier reported findings [23] and used this model to generate *de novo S. cerevisiae* 5'UTR sequences with predictive outcomes in different contexts of promoters, 5'UTRs and protein coding sequences. The broad applicability and predictive power of our model was demonstrated by *in vivo* measurements of fluorescent reporters under diverse translational control conditions. More specifically, R^2 equal to 0.70, 0.73 and 0.67 were obtained for the forward engineered UTR libraries in the RPL8Ap-5'UTR_*RPL8A*-yECitrine, RPL8Ap-5'UTR_*RPL8A*-mTFP1 and TEF1coreP-5'UTR_*TEF1*-yECitrine context respectively. Additionally, the model showed good performance in designing 5'UTRs with predictable behavior for the p-coumaric acid production gene *RcTal1*. Since p-coumaric acid is an important precursor of interesting secondary metabolites, this approach can be useful for future pathway balancing in yeast.

It was also shown that it is important to use the correct coding sequence when calculating 5'UTR strength in yeast, especially due to the possible formation of secondary structures between the 5'UTR sequence and the coding sequence which have a strong influence on translation efficiency. Doing so, it is also possible to estimate the *in vivo* behavior of existing 5'UTRs by reverse engineering. In contrast to what was expected, no increase in protein expression was observed when only replacing the *RPL8A* promoter by the strong *TEF1* core promoter suggesting a strong coherence between transcription and translation. A hypothesis which is strengthened by the fact that most yeast promoters have different transcription start sites and thus different 5'UTR lengths. As such, the demarcation between transcriptional and translational regulatory elements is far from fixed in yeast. Optimally, eukaryotic models predicting 5'UTR effects could therefore be extended with features that influence protein expression at the core promoter level [196].

The *in vivo* behavior of the novel 5'UTRs could largely be explained by the presence of specific sequence motifs such as several 4-mers, the Kozak sequence, uAUGs/uORFs and the nucleotide at position -3 in front of the start codon, all confirming earlier results. Since the accurate prediction of 5'UTRs leading to low protein abundances was rather low, the

identification of additional features is an interesting future perspective. Yet, the detection of low expressed fluorescent proteins is more challenging due to the loss of signal in background noise. To this end, a 5'UTR library in front of the SunTag fluorescence tagging system able to amplify fluorescent signals [268] can be used to assess the effect of 5'UTRs leading to low protein abundances. In addition, this method can be used to determine the entire *in vivo* translation dynamics of mRNA over time [269]. Together with today's high-throughput approaches for oligo synthesis and the available RNA-seq data analysis options [270] (*e.g.* STAR [271], TopHat [272] and GEM [273]) to identify relevant 5'UTR features, a lot of data linking 5'UTR features with mRNA translation dynamics could be generated, as such expanding our knowledge of the effect of the yeast 5'UTR on translation efficiency.

Altogether, we successfully developed a PLS model able to predictably design *de novo* 5'UTRs and reliably calculate strengths of existing 5'UTRs, both for different contexts of promoters and protein coding sequences. As such, our yUTR calculator expands the palette of existing eukaryotic techniques for reliable pathway engineering, all speeding up the development of microbial cell factories.

CHAPTER 5 CRITICAL EVALUATION OF MULTICISTRONIC GENE EXPRESSION IN *SACCHAROMYCES CEREVISIAE*

5.1 ABSTRACT 89
5.2 INTRODUCTION 90
5.3 MATERIAL AND METHODS 93
5.3.1 Strains and media 93
5.3.2 Construction of vectors for fluorescent protein transcription units 93
5.3.3 Characterization of synthetic T2A derivatives 94
5.3.4 Assessment of increasing consecutive T2A numbers 94
5.3.5 Fluorescence and absorbance measurements 95
5.3.6 Western Blotting 96
5.3.7 Data analysis 97
5.4 RESULTS AND DISCUSSION 98
5.4.1 Expanding the T2A palette 98
5.4.2 Genomic integration of T2A sequences at the URA3 locus 103
5.4.3 Tri – and quadcistronic gene expression at the URA3 locus 106
5.5 CONCLUSION 112

Authors:

Thomas Decoene, Sofie De Maeseneire and Marjan De Mey

Publication status:

Unpublished

Author contributions:

TD, SDM and MDM were involved in the conception, design and writing of the manuscript. All experiments, data analysis and interpretation of the results were performed by TD.

5.1 ABSTRACT

Polycistronic expression in eukaryotic cells using 2A peptides has already proven its value, for example in biomedical research and plant biotechnology. Recently, 2A peptides also led to the successful production of secondary metabolites in different yeast species. However, the lack of a thorough evaluation of 2A peptides in yeast, in bi- or multicistronic constructs, makes their general usage for heterologous pathway fine-tuning limited. In this study, we therefore designed a set of five 2A peptides based on the T2A sequence of the *Thosea asigna* virus and characterized them for gene expression (measured as fluorescence) and splicing efficiency (by Western blotting) in bi-, tri- and quadcistronic constructs. This study was performed with *Saccharomyces cerevisiae* and the constructs were genome-integrated. Four out of five T2A peptides showed good activity. Results also revealed that fluorescence of the reporters decreased with an increasing number of open reading frames in the polycistronic construct. Moreover, fluorescence showed a lot of variability at the third position in quadcistronic transcripts and completely dropped to zero at the last position, which indicates a decline in successful splicing and subsequent translation events of farther positioned coding sequences. The efficiency of the 2A peptides was, in some cases, position dependent. In conclusion, bi – or tricistronic expression in *S. cerevisiae* is feasible, yet higher cistron numbers should be omitted. This study can serve as a basis for the application of 2A peptides in pathway engineering for the development of eukaryotic microbial cell factories.

5.2 INTRODUCTION

Transforming microorganisms into powerful microbial cell factories able to efficiently produce specialty metabolites requires the introduction of non-native biosynthetic pathways derived from plants, fungi or other natural sources. While technological methods for heterologous pathway assembly are amply available [8–13], larger difficulties at the moment are found in the fine-tuning of pathways in their new host. Since the metabolic network is tightly regulated and enzyme levels are adjusted for optimal growth, the insertion of novel biosynthetic pathways can have drastic effects on this well-balanced cellular environment. It is therefore of an inordinate importance to tune all steps in these pathways and find a balance with the native cellular background. Despite excellent current know-how and documented successes [5,7,34,79], balancing a new pathway in a host organism to ensure economically viable titers is still one of the main challenges in industrial biotechnology. Especially in the eukaryotic model organism *S. cerevisiae*, where the number of available tools for pathway optimization is rather limited compared to the bacterial *E. coli* model, this tuning process can be a tough exercise.

One strategy for pathway fine-tuning is modifying gene expression at the translational level. In *E. coli*, the ribosome binding site (RBS) is generally well determined and *in silico* software to predictably modify translation initiation rates is accessible [30,179–181]. In contrast, no fixed consensus sequence for ribosomal binding is described for eukaryotes [235] and they have multiple upstream open reading frames (uORFs) in their 5' untranslated regions (5'UTR) that influence translation initiation [251]. Hence, despite some recent breakthroughs [23,239] to which this Ph.D. thesis also contributed [274] (Chapter 4), gene expression levels in *S. cerevisiae* are commonly still tuned at the transcriptional level, where the currently available promoters are seen as an upstream sequence before the start codon without consideration of the 5'UTR. Modulating the transcriptional level of genes is typically realized by using promoters and terminators with a broad range of different strengths [110,188,191,192,202]. In yeast however, the collection size is rather small and mainly exists of native transcriptional control elements. Adding the fact that every coding sequence (CDS) in eukaryotes needs a promoter and terminator, balancing a large biosynthetic pathway in *S. cerevisiae* requires the repeated use of promoters and terminators. This unfortunately increases the chance for homologous recombination and thus risk on strain instability.

The current difficulties caused by a repetitive use of promoters and terminators can be encountered by two strategies. First, new-to-nature synthetic regulatory elements can be developed (Chapter 3) which reduce problems with homologous recombination and at the same time unwanted cellular interactions [24,111,112]. However, the vague definition and the large size of eukaryotic promoters and terminators makes their development by randomization a cumbersome exercise, rapidly resulting in millions of candidates. Furthermore, this combinatorial explosion requires expensive high-throughput screening machines which are not accessible in every lab. A second option is to reduce the number of regulatory elements required by mimicking bacterial polycistronic expression. Currently, two main biological systems exist to establish eukaryotic polycistronic expression. Internal Ribosomal Entry Sites (IRES) are sequences within the mRNA where the ribosome can initiate translation independently of the 5' cap structure. Despite their potential for bicistronic expression, these IRES elements have some major limitations such as their large size (around 500 bp), which complicates cloning work and also causes homologous recombination risks, and their low effectiveness [275,276]. For example, it was proven that the expression of the downstream gene could be as 10 times lower than the upstream gene [277]. A valuable alternative for IRES are 2A peptides, short peptides of up to 20 amino acids obtained from viral polyproteins [277–285]. Well-known examples are F2A, originating from the foot-and-mouth disease virus, P2A, originating from porcine teschovirus-1, and T2A, originating from the *Thosea asigna* virus. 2A peptides contain a conserved 'NPGP' sequence at their C-terminus where the ribosome skips the linkage between the glycine and proline, leading to an upstream protein with a short C-terminal 2A peptide tag and a downstream protein with a proline at its N-terminus. As a result, by placing 2A peptide sequences between the pathway CDSs, multiple separate proteins can be produced from a single open reading frame (ORF). In addition, 2A peptides have a good splicing efficiency yielding equimolar amounts of proteins and have a short nucleotide sequence (60-70bp) [277], ideal to be used in primers and as linkers for *in vivo* recombination.

The potential of 2A peptides for polycistronic pathway expression has already been assessed in yeasts such as *Pichia pastoris* for the production of β-carotene [286], violacein [286] and glycine betaine [287], and *S. cerevisiae* for the production of β-carotene and β-ionone [288], serotonin derivatives [289] and triterpene saponins [35]. In the latter three studies performed in *S. cerevisiae*, no more than three CDSs were expressed from one ORF and expression vectors were used instead of integrating the transcription units into the genome, though the latter

is highly recommended to ensure strain robustness. As *S. cerevisiae* is also gaining interest as an industrial cell factory for specialty metabolites produced via large biosynthetic pathways, it is essential to explore the maximum number of CDSs that can be efficiently expressed in a single ORF as this might allow building larger and more stable pathways using these promising 2A peptides. Therefore, the objective of this study was to investigate to what level higher numbers of consecutive 2A peptides are efficiently spliced and if subsequently, adequate protein activities were observed. We used specifically the 2A peptide sequence from *Thosea asigna* virus (T2A) as its functionality was already confirmed in *S. cerevisiae* [288] and *P. pastoris* [286]. The constructs were inserted at the *URA3* locus. Furthermore, we modified the existing T2A peptide sequences to avoid homologous recombination between multiple T2As in the genome. As a result, a new set of characterized 2A peptides becomes available for polycistronic pathway expression in *S. cerevisiae*.

5.3 MATERIAL AND METHODS

All products were purchased by Sigma-Aldrich (Diegem, Belgium) unless otherwise stated. All DNA fragments for Circular Polymerase Extension Cloning (CPEC) [11] and genomic integration were amplified using PrimeSTAR HS DNA polymerase (Takara, Westburg, Leusden, The Netherlands) and purified using the innuPREP PCRpure Kit (Analytik Jena AG, Jena, Germany). All plasmids were isolated from bacterial cultures using the innuPREP Plasmid Mini Kit (Analytik Jena AG).

5.3.1 Strains and media

S. cerevisiae strain BY4742 (*Matα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*, Euroscarf, University of Frankfurt, Germany [205]) was used in this study as expression host. Yeast strains were grown in synthetic defined (SD) medium consisting of 0.67% YNB without amino acids, 2% glucose (Cargill, Sas van Gent, The Netherlands) and selective amino acid supplement mixture (MP Biomedicals, Brussels, Belgium) dependent on the required auxotrophies. For solid media, an extra 2% Agar Noble (Difco, Erembodegem, Belgium) was added. One Shot TOP10 Electrocomp™ *E. coli* (ThermoFisher Scientific, Aalst, Belgium) was used for cloning procedures and for maintaining plasmids. *E. coli* strains were cultured in lysogeny broth (LB) consisting of 1% tryptone-peptone (Difco), 0.5% yeast extract (Difco) and 0.5% sodium chloride (VWR, Leuven, Belgium). For solid LB growth medium, 1% agar (Biokar diagnostics, Pantin Cedex, France) was added. All strains used in this study are listed in Supplementary Table S.3.1.

5.3.2 Construction of vectors for fluorescent protein transcription units

Four plasmids with a monocistronic transcription unit (TU) expressing a fluorescent reporter protein (yECitrine, mCherry, mTagBFP2 or mTFP1) under control of the *TEF1* promoter [159] and *ADH1* terminator [206] were developed. Both promoter and terminator sequence were PCR-amplified from genomic DNA of *S. cerevisiae* BY4742 (Genbank JRIR00000000). The reporter coding sequences for yECitrine, mCherry and mTagBFP2 were picked up from pKT140 (Euroscarf [206]), iGEM part BBa_J06504 and iGEM part BBa_K592100 containing an Ile174Ala amino acid replacement, respectively. The mTFP1 sequence was ordered as a gBlock (Genbank DQ676819 [240], Integrated DNA Technologies (IDT, Leuven, Belgium)). The four TUs were assembled in an in-house *CEN6/ARS4* low copy backbone with the *URA3* auxotrophic marker (p2a backbone), as such creating the

fluorescent protein (FP) TU vectors p2a33_yECitrine, p2a33_mCherry, p2a33_mTagBFP2 and p2a33_mTFP1.

5.3.3 Characterization of synthetic T2A derivatives

To characterize new T2A sequences low copy T2A characterization plasmids carrying bicistronic constructs of yECitrine and mCherry were constructed using CPEC [11] (Figure 5.1B.). The yECitrine and mCherry carrier vectors were used as templates to pick-up the BB_a-pTEF1-yECitrine and mCherry-tADH1-BB_b part, respectively (BB_a and BB_b are respectively both halves of the p2a backbone). In all designs, CDSs preceding a T2A were lacking a stop codon and CDSs following a T2A missed the start codon. For the introduction of the T2A peptide sequence between two FPs, fragments were PCR-amplified using 80 bp primers including the 60 bp T2A sequence serving as homologous overlap for CPEC. The bicistronic T2A characterization constructs were expressed in the p2a backbone and were verified by colony PCR using *Taq* DNA polymerase (NEB, Bioké, Leiden, The Netherlands) and sequencing (EZ-Seq, Macrogen, Amsterdam, The Netherlands). The negative control vector p2a_empty in sRepb consists of the p2a backbone without promoter, CDS and terminator.

Yeast transformations in strain BY4742 were carried out using the lithium acetate method [208]. After transformation, strains were incubated on selective SD medium at 30°C for 3-4 days. Correct *S. cerevisiae* strains were confirmed by colony PCR using OneTaq 2X Master Mix with Standard Buffer (NEB). All plasmids used and constructed are listed in Supplementary Table S.3.2.

5.3.4 Assessment of increasing consecutive T2A numbers

To assess the splicing efficiency of increasing numbers of consecutive T2A sequences, bi-, tri – and quadcistronic constructs were integrated in the genome at the *URA3* locus using a combination of CRISPR/Cas9 and *in vivo* assembly (Figure 5.3). To construct a Cas9 expressing yeast strain, the *TRP1* auxotrophic marker of p414-pTEF1-Cas9-CYC1t (Addgene #43802 [19]) was replaced by the *LEU2* marker of plasmid p415-GalL-Cas9-CYC1t (Addgene #43804 [19]) by two-pieces CPEC [11]. The resulting pCas9L plasmid was transformed in BY4742 by the lithium-acetate method [208] and led to strain sCas9L. For integration of the DNA fragments at the *URA3* locus, a gRNA expression plasmid (p_gRNA_URA3) was constructed. First, a template plasmid, p426-SNR52p-*aeBlue*-SUP4t, was made by replacing

the original gRNA sequence of p426-SNR52p-gRNA.CAN1.Y-SUP4t (Addgene #43803 [19]) with *aeBlue* (iGEM part BBa_K864401). This allows easy selection of correct clones after CPEC in *E. coli*. This vector was then used as template to amplify the gRNA expression backbone. A gRNA targeting the *URA3* locus was chosen based on the *URA3* knock-out sequence from Brachmann *et al.* [205]. The gRNA sequence (5' tcagggtccataaagctccc 3') was ordered as a 60bp oligonucleotide (IDT) where the 20bp gRNA sequence was flanked at each side with 20bp compatible backbone ends for CPEC. For the production of DNA fragments for genomic integration, 500 bp up – and downstream homologies were PCR-amplified from *S. cerevisiae* BY4742 genomic DNA (Chromosome V from 115642 to 116166 and from 116971 to 117501). The 500 bp upstream homology contained the SHR_G linker sequence [290] at its 3' end and the downstream the SHR_F linker [290] at its 5' end. These linkers served as 60 bp homologous overlap for *in vivo* assembly with the *TEF1* promoter and *ADH1* terminator, respectively. FP fragments were PCR-amplified from their TU vectors using 80 bp primers including the 60 bp T2A sequence serving as homologous overlap. Also here, CDSs preceding a T2A were lacking a stop codon and CDSs following a T2A missed the start codon. For *in vivo* assembly and integration of the different T2A sequence combinations at the *URA3* locus, 0.4 picomoles of each DNA fragment together with 1µg of gRNA vector were transformed in sCas9L by the lithium-acetate method [208]. Correct colonies were verified by colony PCR using OneTaq 2X Master Mix with Standard Buffer (NEB). All strains are listed in Supplementary Table S.3.1.

5.3.5 Fluorescence and absorbance measurements

For fluorescence and absorbance measurements, yeast strains were first grown in sterile 96-well flat-bottomed, black microtiter plates (Greiner Bio-One, Vilvoorde, Belgium) enclosed by a Breath-Easy® sealing membrane (Sigma-Aldrich) containing 150 µl selective SD medium. For every experiment, three biological replicates were inoculated from agar plate and incubated on a Compact Digital Microplate Shaker (ThermoFisher Scientific, 3mm orbit) at 800 rpm and 30°C for 24h. Subsequently, these pre-cultures were diluted 150 times in 150 µl fresh selective SD medium. After 26h of growth (stationary phase), optical density (OD) and fluorescence (FP) of the *S. cerevisiae* strains was measured using the TECAN Infinite® 200 PRO (Tecan). OD of yeast cultures was measured at 600 nm, excitation wavelengths of mTagBFP2, mTFP1, yECitrine and mCherry were respectively 415 nm, 460

nm, 500 nm and 575 nm. Emission of mTagBFP2, mTFP1, yECitrine and mCherry was measured at 460 nm, 500 nm, 540 nm and 620 nm respectively.

For analysis of fluorescence measurements, two types of controls were included on every single MTP. A medium blank (*i.e.* selective SD medium) was used for the correction of background absorbance of the medium (OD_{bg}). sRepb and sCas9L containing respectively p2a_empty and pCas9L were used to correct for the background fluorescence of yeast cells (FP_{bg}). sRepb served as control for the T2A plasmid expression strains while sCas9L was used as control for the yeast strains expressing T2A constructs integrated in the genome. For all strains, OD corrected fluorescence was used as measure for fluorescent gene expression and calculated as follows:

$$\left(\frac{FP}{OD}\right)_{corrected} = \frac{FP - FP_{bg}}{OD - OD_{bg}} \tag{5.1}$$

The relative fluorescence was defined as follows:

$$Relative\ fluorescence\ (\%) = \frac{\left(\frac{FP}{OD}\right)_{corrected}}{\left(\frac{FP}{OD}\right)_{corrected,\ Ref}} \times 100 \tag{5.2}$$

5.3.6 Western Blotting

Cultures for total protein extraction were grown in 5 ml selective SD medium until stationary phase was reached (OD 5-6). Subsequently, the culture broth was centrifuged and 100 µl CelLytic Y with 1 µl Protease Inhibitor Cocktail was added to the cell pellet and shaken for 30 min at room temperature. Finally, cells were centrifuged at 12000 rpm to collect supernatant.

The total protein concentration was determined by the Pierce™ BCA Protein Assay Kit (ThermoFisher scientific) and 10 µg of total protein was used for sample preparation with 20µl Laemmli Sample Buffer supplemented with 5% 2-mercapto-ethanol. The Laemmli Sample Buffer was composed of 62.5 mM Tris-HCl, pH 6.8 (Promega Benelux, Leiden, The Netherlands), 25% glycerol (Chem-Lab Analytical, Zedelgem, Belgium), 2% SDS and 0.01% bromophenol blue. After heating at 95°C for 15 min, 10 µl of samples were loaded on 12%

SDS-PAGE gel together with 3 µl PageRuler Prestained Protein Ladder (ThermoFisher scientific). The gel was run at 200V (Mini-PROTEAN® System, Bio-Rad, Temse, Belgium) for around 1h and afterwards, proteins were transferred (Mini Trans-Blot®, Bio-Rad) to a nitrocellulose membrane (GE Healthcare Life Sciences, Diegem, Belgium) by blotting in CAPS buffer (10 mM CAPS, pH 11, 10% methanol) during 1h at 100V. Membranes were blocked overnight in Phosphate Buffered Saline (PBS) buffer (100 mM NaCl, 33 mM $Na_2HPO_4.2H_2O$ and 17 mM NaH_2PO_4) with 1% casein. After washing three times with PBS containing 0.2% Triton X100, the membranes were incubated for 2h with primary antibodies anti-GFP (mouse), anti-mCherry (mouse, Clontech, Westburg, Leusden, The Netherlands) or anti-2A (rabbit). Again, the membranes were washed three times with PBS containing 0.2% Triton X100, and incubated now for 1h with alkaline phosphatase secondary anti-mouse or anti-rabbit antibodies. Finally, membranes were washed with PBS buffer and proteins of interest were visualized using a colorimetric alkaline phosphatase system. Therefore, the membranes were incubated in the dark for 30 min at 37°C in 10 ml phosphatase buffer (10 mM Tris pH 9.5, 100 mM NaCl, 50 mM $MgCl_2$) supplemented with 50 µl nitroblue tetrazolium/5-bromo-4-chloro-3-indolyl phosphate (NBT/BCIP) stock solution. Protein bands became visible as a purple-blue colored precipitate and pictures were taken with a Gel-Doc™ XR+ Gel Documentation System (Bio-Rad). Band intensity analysis and quantification of the Western blots was performed by ImageJ version 1.51J8. Splicing efficiency was calculated as follows:

$$Splicing\ efficiency = \frac{spliced\ protein}{spliced\ protein + unspliced\ protein} \tag{5.3}$$

5.3.7 Data analysis

All calculations were performed in Python using the Python Data Analysis Library (Pandas). Error bars represent the standard error of the mean (n = 3). Pairwise comparisons between different strains were done by a two-sided T-test using the scipy.stats package in Python. In all cases, a significance level of 0.05 was applied.

5.4 RESULTS AND DISCUSSION

5.4.1 Expanding the T2A palette

To avoid homologous recombination between repeatedly used 2A-encoding sequences in polycistronic pathways, T2A peptides that differ as much as possible in nucleotide sequence are crucial. Hence, as a first step, the palette of T2A peptides available for *S. cerevisiae* was extended. In *P. pastoris,* up to nine genes were successfully expressed using eight altered T2As by Geier *et al.* [286]. The sequences used in that research served here as templates for the development of new, efficient T2A peptide coding sequences for *S. cerevisiae*. To this end, all eight T2A sequences were first extended with a glycine-serine-glycine (GSG) tag at their N-terminus to improve cleavage efficiency [280] and additionally, the first four 5'-nucleotides and last four 3'-nucleotides were kept the same for possible future applications as linkers in Golden Gate based pathway assemblies such as VEGAS [21]. After a multiple sequence alignment (Supplementary Figure S.3.1), Geier's modified sequences T2A1*, T2A2* and T2A6* were selected since T2A2* and T2A6* did not show more than 75% nucleotide identity with T2A1*. As the homologous recombination activity in *S. cerevisiae* is rather high and consequently too much similarities must be avoided, only these T2A-encoding sequences were used for further modification. T2A1*, T2A2* and T2A6* were renamed as T2A1, T2A2 and T2A3, respectively (Figure 5.1A). Next, extra alterations were introduced to strive toward a common nucleotide difference of 30% or more, since it was proven that recombination then no longer occurs [288]. Alterations were added as silent mutations in the three sequences, except for one amino acid replacement from glutamate to serine on position 17 (S17E) in T2A3. Specifically this amino acid position was targeted to increase the sequence difference above the 30% target since it was shown that substitutions here had little influence on 2A peptide cleavage activity [291]. As a result, a nucleotide identity of 68% was achieved for T2A1 with T2A2 and T2A3, and an identity of 65% was reached between T2A2 and T2A3 (Supplementary Figure S.3.1). Furthermore, to further expand the T2A palette, two novel 2A-encoding sequences based on the combination of an F2A and T2A peptide were designed and evaluated. The much less conserved N-terminus of T2A was combined with the strongly conserved C-terminus of F2A and vice versa, yielding two chimeric 2A sequences, T2Ac1 and T2Ac2 (Figure 5.1A). With the introduction of extra silent mutations, a sequence identity of only 50% was reached between T2Ac1 and T2Ac2. A final multiple sequence alignment between these five novel T2A peptide sequences,

showed that only T2A1 and T2Ac1, and T2A3 and T2Ac1 had more than 70% nucleotide identity, while all the others reached the suggested threshold of more than 30% nucleotide difference (Supplementary Figure S.3.1).

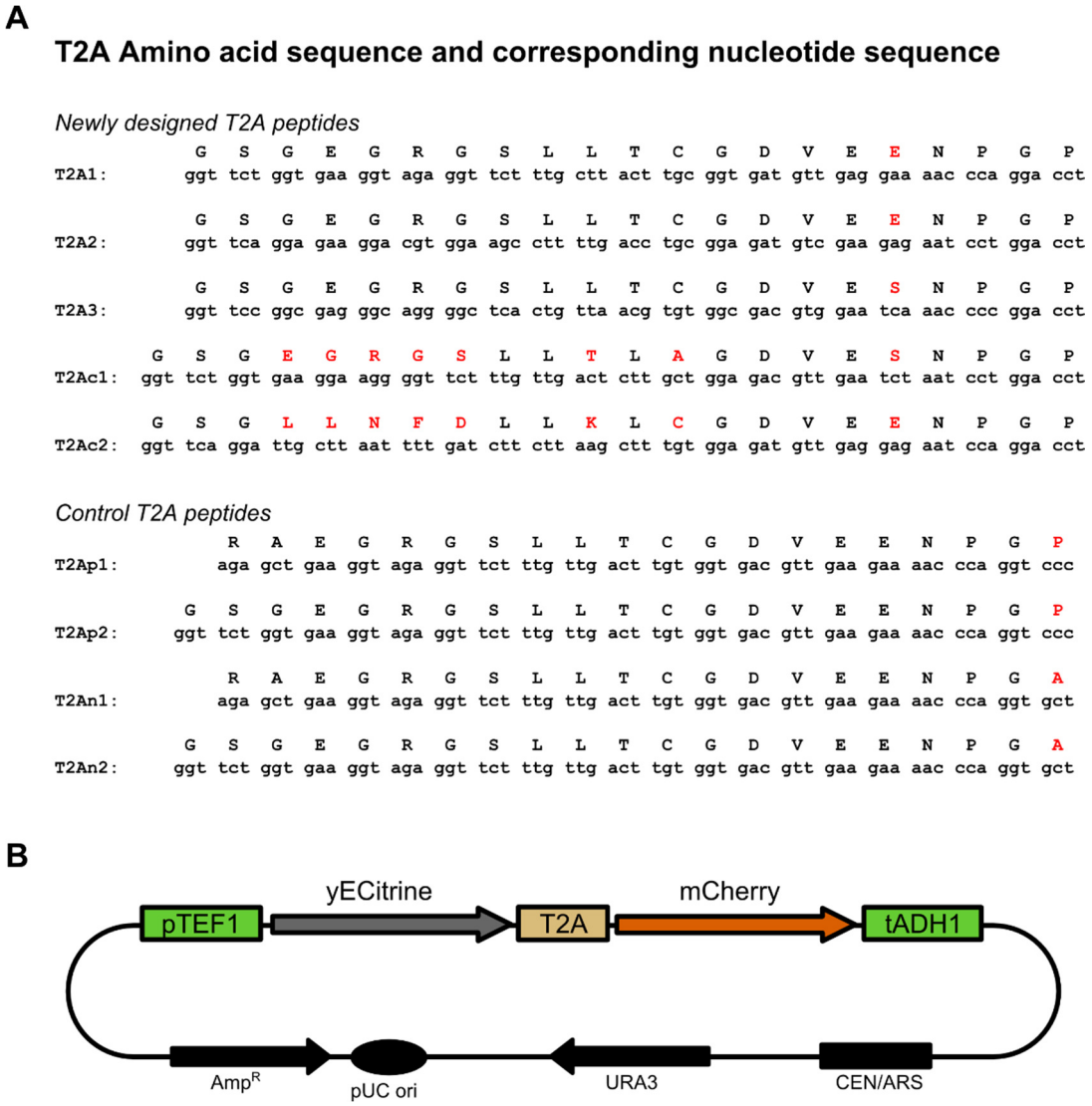

Figure 5.1: (A) Amino acid and nucleotide sequences of the T2A peptides characterized in *S. cerevisiae* (T2A1, T2A2, T2A3, T2Ac1 & T2Ac2), including two positive (T2Ap1 & T2Ap2) and negative controls (T2An1 & T2An2). (B) Schematic representation of the yeast expression vector used for the characterization of the aforementioned T2A peptides (Supplementary Table S.3.2). The yeast *TEF1* promoter (pTEF1) and *ADH1* terminator (tADH1) are indicated in green, marker genes and origins of replication are colored black (respectively, AmpR & pUC ori for *E. coli* and *URA3* & CEN6/ARS4 for *S. cerevisiae*).

The activity in *S. cerevisiae* of each of these five T2A sequences was tested using a bicistronic construct in which yECitrine and mCherry were separated by the designed T2As (Figure

5.1B, Supplementary Table S.3.2). Since it was shown here (Supplementary Figure S.3.2) and in literature [286] that gene expression levels are not influenced by the position relative to the T2A peptide in a bicistronic TU, yECitrine was always placed in front of mCherry. T2Ap1 without GSG-tag and T2Ap2 with GSG-tag served as positive controls and were based on the original T2A-encoding sequence from Beekwilder *et al.* [288]. For the negative controls, the essential amino acid for ribosomal skipping proline was replaced in T2Ap1 and T2Ap2 by an alanine, yielding defective T2A peptides T2An1 and T2An2 (Figure 5.1A).

As such, nine *S. cerevisiae* strains were constructed (Supplementary Table S.3.1) and gene expression and splicing efficiency of the T2As was examined by measuring fluorescence (Figure 5.2A) and by Western blotting (Figure 5.2B and C). In all strains, yECitrine fluorescence was reduced by half compared to the reference strain expressing the single protein. This confirmed results described in literature [286,292] and of experiments with yEGFP performed in this study (Supplementary Figure S.3.2). The decrease in expression could be explained by the fact that in a same period of time, only half of the T2A construct mRNAs are formed compared to the monocistronic reference, since the transcripts are doubled in length (future qPCR experiments should be performed to confirm this hypothesis). In eukaryotes, transcription and translation is separated in space and time [293] and thus no simultaneous transcription and translation occurs. As such, a reduced number of reporter mRNAs would lead to lower fluorescence levels. However, except for T2Ac1, mCherry fluorescence was significantly higher than with the monocistronic reference (all p-values at least < 0.03, Figure 5.2A). This is remarkable, since, to the best of our knowledge, only a decrease in gene expression at the second position of a bicistronic construct compared to its monocistronic counterpart has been described [292,294] and detected in our own experiments (Supplementary Figure S.3.2). As an extra control, the positional influence of the mCherry reporter was evaluated by switching its position with yECitrine. Again, the position had no influence on the mCherry fluorescence in the bicistronic TU since also in this construct mCherry fluorescence was higher (Supplementary Figure S.3.3). In a second control, quenching of the fluorescent reporters was investigated by measuring fluorescence of mixed protein extracts of strains sRep1 and sRep2, but no increased mCherry fluorescence was observed for the mixed samples compared to the single mCherry reference (data not shown).

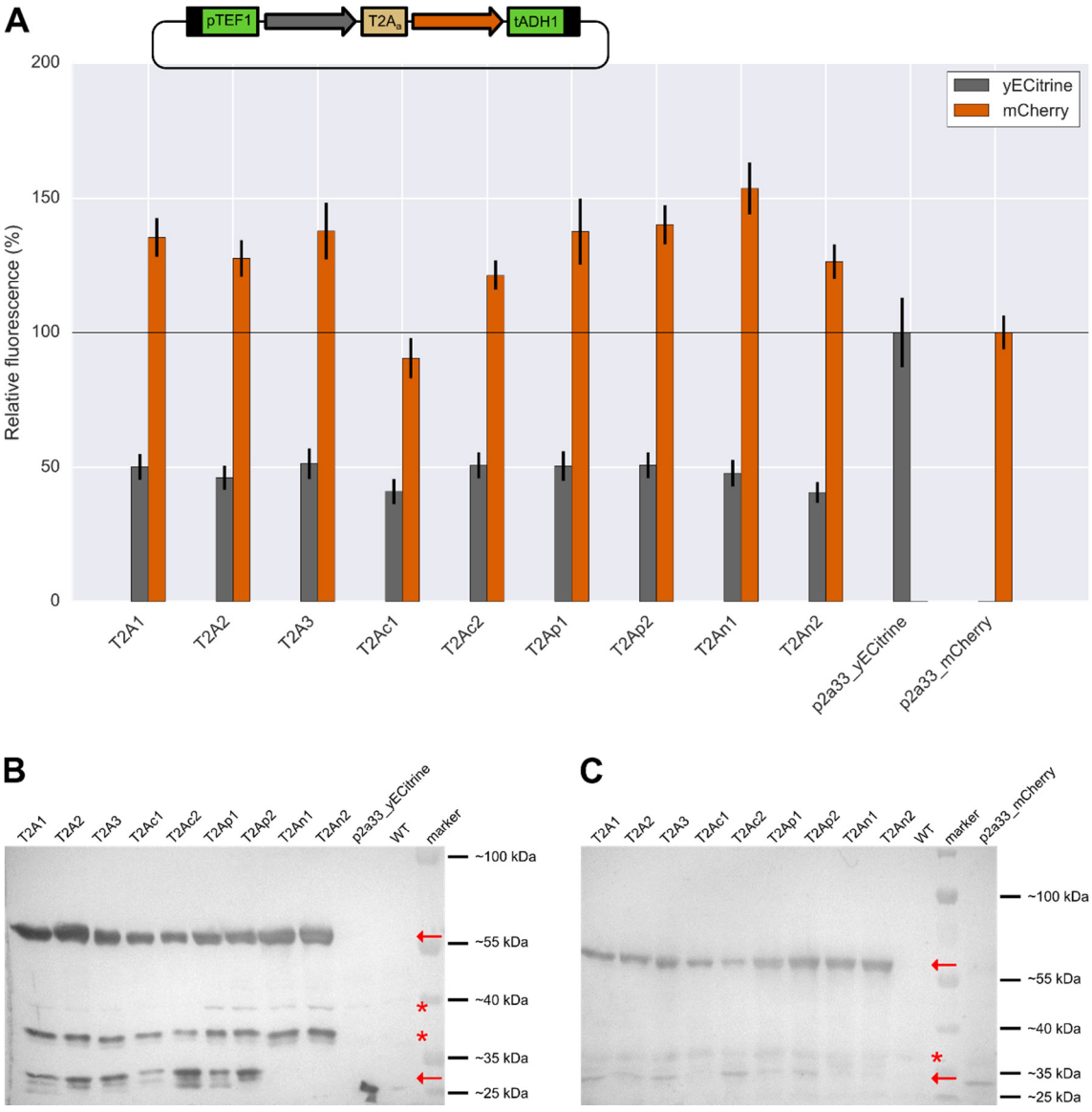

Figure 5.2: Characterization of T2A peptides on a low copy, bicistronic expression vector using yECitrine and mCherry. The strains corresponding with the represented T2A peptides are listed in Supplementary Table S.3.1. (A) Fluorescence of yECitrine and mCherry as a measure for gene expression was normalized to their monocistronic reference strains (represented by the horizontal line). Error bars represent the standard error of the mean of three biological replicates. (B) Western blot for yECitrine detection using anti-GFP. (C) Western blot for mCherry detection using anti-mCherry. Red arrows represent cleaved (bottom) and uncleaved (top) protein products, asterisks indicate unknown detected byproducts.

As such, it still needs to be explained why mCherry gave rise to higher fluorescence levels, although intrinsic properties of the reporter proteins presumably lay at the basis of this result since the effect was for example not seen when yEGFP was placed at the second

position (Supplementary Figure S.3.2). The negative control strains showed fluorescence levels in the same order of magnitude of the positive controls, indicating that also the uncleaved proteins highly fluoresce. This is not unlikely since fluorescent reporters are used a lot as protein localization tags [206,295]. Hence, it is suggested that the measured marker fluorescence corresponds with the total spliced and unspliced protein activity. To this end, the observed fluorescence levels can not be used to reveal anything about the separate activities of cleaved and uncleaved proteins and are in addition unusable to observe T2A splicing performance. Therefore, Western blotting was carried out to analyze the cleavage efficiency of the novel T2A peptides.

Total protein extracts were used to determine each T2A peptide's splicing efficiency by Western blot. Three different antibodies were used (anti-GFP, anti-mCherry and anti-2A) to investigate if the fluorescent reporters are present as separate proteins (~ 28 kDa) or as fusion protein (~ 55 kDa). Western blot results are given in Figure 5.2B and C, and Supplementary Figure S.3.4. A detailed overview of all possible protein products is given in Supplementary Table S.3.3. For all five T2A sequences under study (T2A1 to 3, T2Ac1 and T2Ac2) bands for the spliced proteins were visible with different antibodies, which illustrates cleavage activity of these T2A peptides. However, T2Ac1 seems to have less efficient cleavage capacity compared to the others as a lighter band is present for anti-GFP at 28 kDa and no band appeared with anti-mCherry. This could be explained by successful skipping of the ribosome and release of the first protein, but ribosome fall-off and discontinued translation of the second protein [294]. Also the previously reported positive influence on splicing efficiency of the GSG-tag [280,285,294,296] was illustrated with T2Ap1 (without GSG) and T2Ap2 (with GSG) leading to respectively lighter and slight darker bands with anti-GFP and anti-mCherry (Figure 5.2B and C, Supplementary Figure S.3.5). In addition, byproducts were detected on both Western blots (Figure 5.2B and C, indicated with an asterisk). Possibly, these bands correspond to protein degradation products of the T2A bicistronic constructs caused by remaining protease activity in the raw protein extracts.

Unfortunately, an extensive amount of uncleaved fusion products of yECitrine and mCherry, inclusive for the positive controls, was also detected on the immunoblots, demonstrating that the splicing efficiency of T2A peptides in our BY4742 yeast strain is never 100%, *i.e.* splicing efficiencies for T2A sequences on plasmid were never higher than 50% and

generally around 20% (Supplementary Figure S.3.5). This is in contrast to the results obtained by Beekwilder *et al.* [288] with the CEN.PK yeast strain and by Wang *et al.* [285] with insect cell lines, where for the latter T2A cleavage efficiencies of 90-97% were described. On the other hand, our results are similar to what was seen in *P. pastoris* [286] in that way that obvious uncleaved protein bands were visible on Western blot, and to observations in human cell lines and mice [280] where T2A splicing efficiency was far lower compared to their P2A counterpart. While some studies claim that P2A peptides are the most efficient and others state that T2A peptides are better, these results indicate that cleavage efficiency is strongly host-dependent, even amongst different yeast species, and thus prior characterization in the host of interest is essential before usage of 2A peptides in biosynthetic pathways.

5.4.2 Genomic integration of T2A sequences at the URA3 locus

Though in all examples of multicistronic pathway engineering using 2A peptides in yeast TUs are expressed from expression vectors [286–289], robust expression of biosynthetic pathways from the genome is desired in industrial microbial cell factories. To investigate feasibility of the latter, T2A constructs were integrated at the *URA3* locus using CRISPR/Cas9 and *in vivo* yeast assembly (Figure 5.3). The *URA3* chromosomal integration site has been characterized as a suitable spot for genomic pathway insertion [120] and CRISPR/Cas9 combined with *in vivo* assembly in yeast is a very efficient way to quickly and reliably integrate heterologous pathways in one step without the need of auxotrophic markers [297,298].

To start with, fluorescence of four fluorescent reporter proteins (*i.e.* yECitrine, mCherry, mTFP1 and mTagBFP2, see also 5.4.3) was compared between the *URA3* locus and low copy CEN6/ARS4 vectors (p2a backbone). As shown in literature [192], genomic integration of the reporters at the *URA3* locus led to lower variability and also to lower OD corrected fluorescence compared to expression on vector counterparts. More specifically, fluorescence dropped 2.8 to 7-fold when FPs were expressed from the *URA3* locus (Supplementary Figure S.3.6). Results in Supplementary Figure S.3.6 also indicate that the spectra of the different FPs used in this study do not significantly overlap.

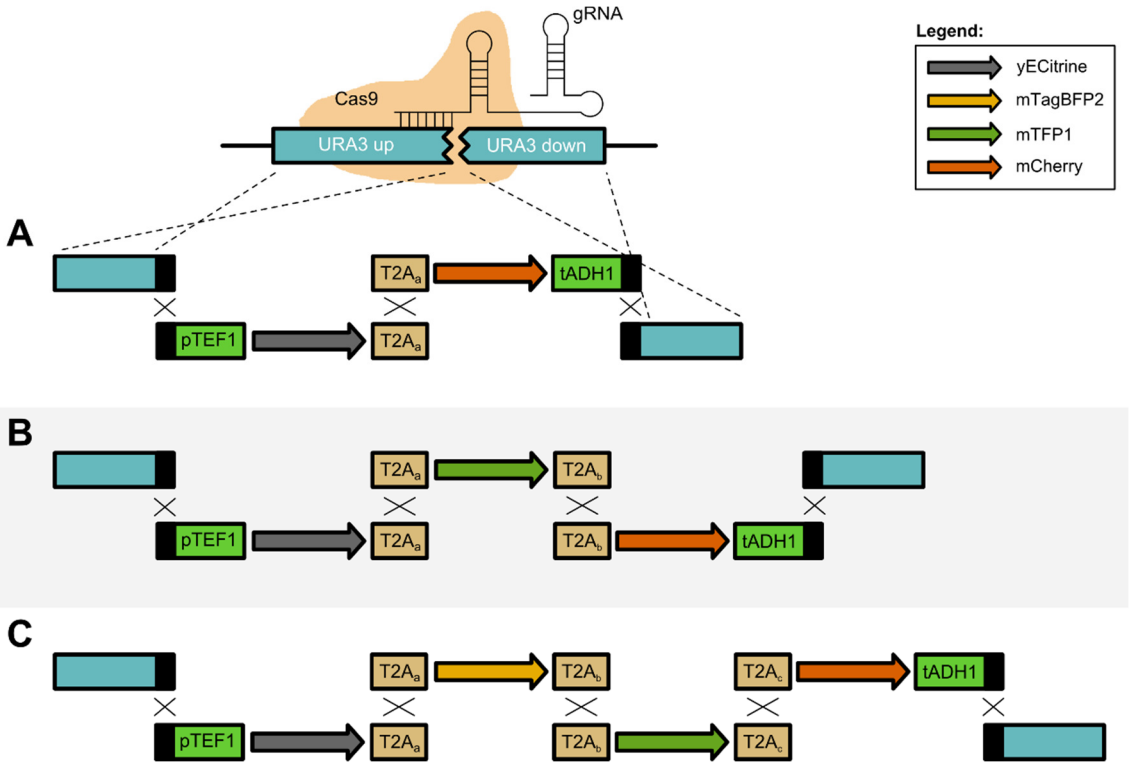

Figure 5.3: Schematic overview of the insertion of T2A expression constructs at the *URA3* locus using *in vivo* assembly and CRISPR/Cas9. The 60 bp T2A peptide sequences served as linkers for *in vivo* homologous recombination between the different reporter proteins. The up – and downstream homologies for the *URA3* locus were around 500 bp in length and contained either the SHR_G or SHR_F linker sequence [290] for homologous recombination with the *TEF1* promoter or *ADH1* terminator sequence (black rectangles). (A) Insertion of a bicistronic T2A expression construct. (B) Insertion of a tricistronic T2A expression construct. (C) Insertion of a quadcistronic T2A expression construct.

Next, again nine *S. cerevisiae* strains were constructed, but now the bicistronic T2A peptide transcription units described in section 5.4.1 were integrated at the *URA3* locus. Fluorescence measurements generally showed the same outcome as for expression from plasmids. Indeed, fluorescence of yECitrine in the bicistronic constructs were 40 to 50% lower than in the yECitrine reference strain (sReg1) and mCherry fluorescence increased up to 150% (Figure 5.4A). Also the Western blots gave similar results as with the plasmid based expression system (Figure 5.4B and C), with the exception however of T2A3. For this sample, the total protein concentration loaded on SDS gel was not high enough to detect yECitrine with anti-GFP (Figure 5.4B), tough the spliced mCherry reporter was detected with anti-mCherry (Figure 5.4C). The unreliable splicing activity of T2Ac1 was also confirmed since no band linked to the single mCherry protein was visible on Western blot

(Figure 5.4C, Supplementary Figure S.3.5). Again, no 100% splicing was achieved, as fused fluorescent reporters remained present (Figure 5.4B and C and Supplementary Figure S.3.5), and a lot of unknown protein byproducts were detected (asterisks in Figure 5.4B and C). Splicing efficiencies of the T2As on the genome were slightly higher compared to the plasmid based expression system (Supplementary Figure S.3.5).

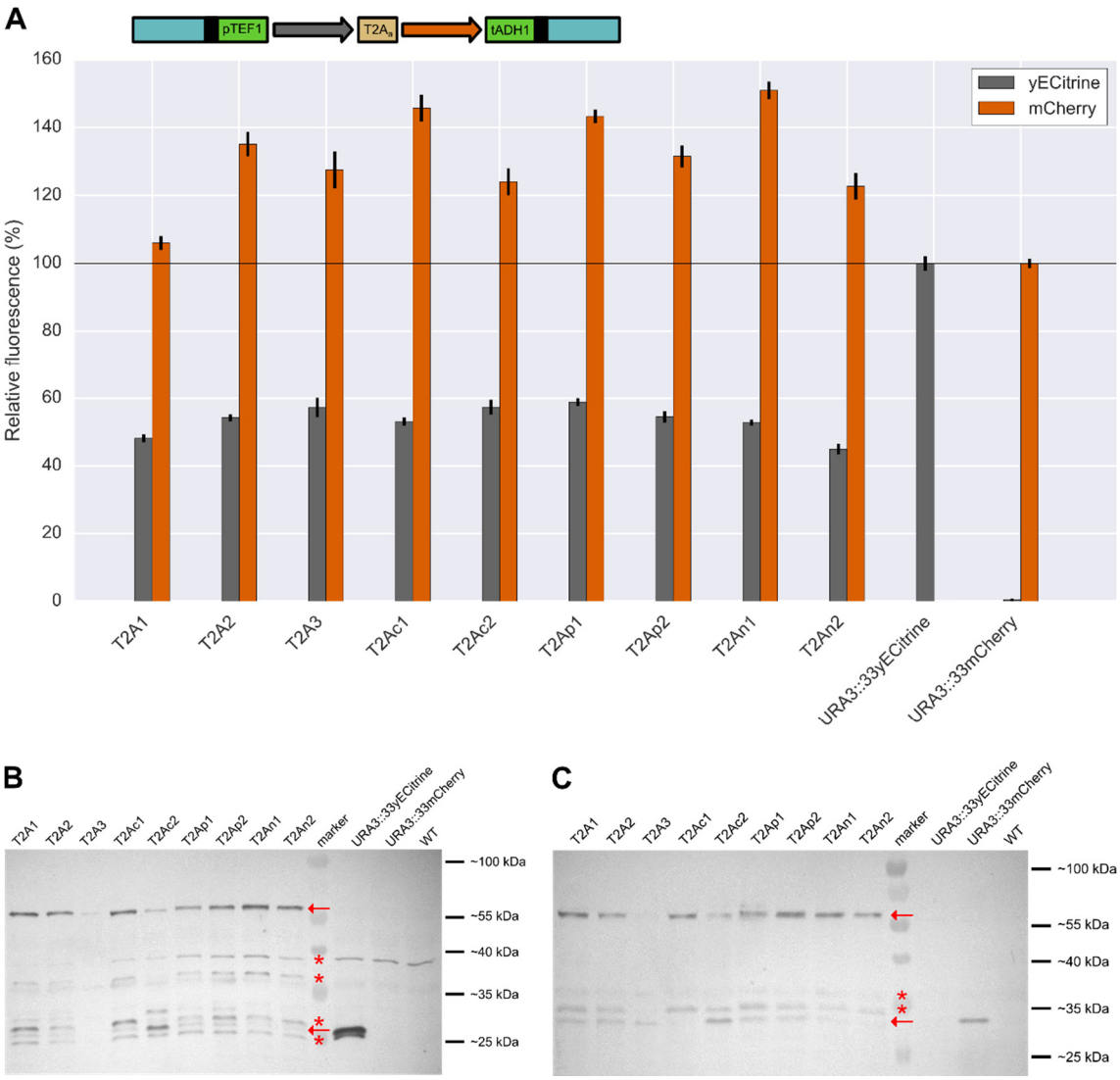

Figure 5.4: Characterization of T2A peptides in a bicistronic construct at the *URA3* locus using yECitrine and mCherry. The strains corresponding with the represented T2A peptides are listed in Supplementary Table S.3.1. (A) Fluorescence of yECitrine and mCherry as a measure for gene expression was normalized to their monocistronic reference strains (represented by the horizontal line). Error bars represent the standard error of the mean of three biological replicates. (B) Western blot for yECitrine detection using anti-GFP. (C) Western blot for mCherry detection using anti-mCherry. Red arrows represent cleaved (bottom) and uncleaved (top) protein products respectively, asterisks indicate unknown detected byproducts.

5.4.3 Tri – and quadcistronic gene expression at the URA3 locus

For the construction of a tricistronic transcription unit under control of the *TEF1* promoter and *ADH1* terminator, three different fluorescent reporters separated by T2A sequences were used (Figure 5.3B). The FPs were incorporated in a fixed order in the yeast genome with yECitrine preceding mTFP1 and mTFP1 preceding mCherry. To evaluate quadcistronic expression in yeast a fourth FP, mTagBFP2, was integrated in the TU at position 2, behind yECitrine and in front of mTFP1 (Figure 5.3C). This design allows to visualize any protein product that can be formed by T2A mediated splicing using anti-2A and anti-mCherry antisera. For example, incomplete processing of the first T2A peptide in a tricistronic transcript would lead to the yECitrine-T2A$_a$-mTFP1-T2A$_b$ fusion product (57.2 kDa) and mCherry (26.3 kDa), detectable by anti-2A and anti-mCherry, respectively (Supplementary Table S.3.3). All ten combinations of the designed T2A sequences possible ($\binom{5}{2}$ and $\binom{5}{3}$ without replacements) in tri – and quadcistronic constructs when fixing the order as described above were tested. Additionally, genomic stability was evaluated since T2A sequences are a potential source for homologous recombination and thus strain instability.

Generally, strain analysis by colony PCR showed no recombination between the different T2A sequences, indicating stable genomic integration for all polycistronic TUs. For tricistronic transcripts, fluorescence measurements were highest at the second and third position, and lowest for the yECitrine reporter at the first position. Specifically, fluorescence dropped with ca. 75% for yECitrine and ca. 40% for both mTFP1 and mCherry compared to monocistronic expression (Figure 5.5A). This is in line with the results of a study in *A. niger* where luciferase as reporter also showed decreased activity at position 1 compared to position 2 and 3 in a tricistronic TU [299]. In contrast to what was observed for mCherry activity in the bicistrons, its fluorescence now was lower in comparison with monocistronically expressed mCherry. Further on, this decreasing trend in fluorescence continued for quadcistronic TUs. yECitrine decreased to ca. 20% of its monocistronic reference strain, yet in this case, the FP at position 2, *i.e.* mTagBFP2, decreased to ca. 25% (Figure 5.6A) in contrast to the observations in the bi- and tricistronic constructs where the FP at position 2 fluoresces much stronger than the FP at position 1. For the FPs at positions 3 and 4 in the quadcistronic constructs, mTFP1 and mCherry, fluorescence even dropped below 15% or was completely eliminated (Figure 5.6A). The overall decrease in fluorescence compared to the bicistronic TUs supports our earlier observations of the

determining role of mRNA length on protein abundance in yeast. Additionally, longer mRNAs are negatively correlated with ribosomal density [300,301] and also a negative correlation between transcript length en mRNA stability in *S. cerevisiae* was observed [302,303].

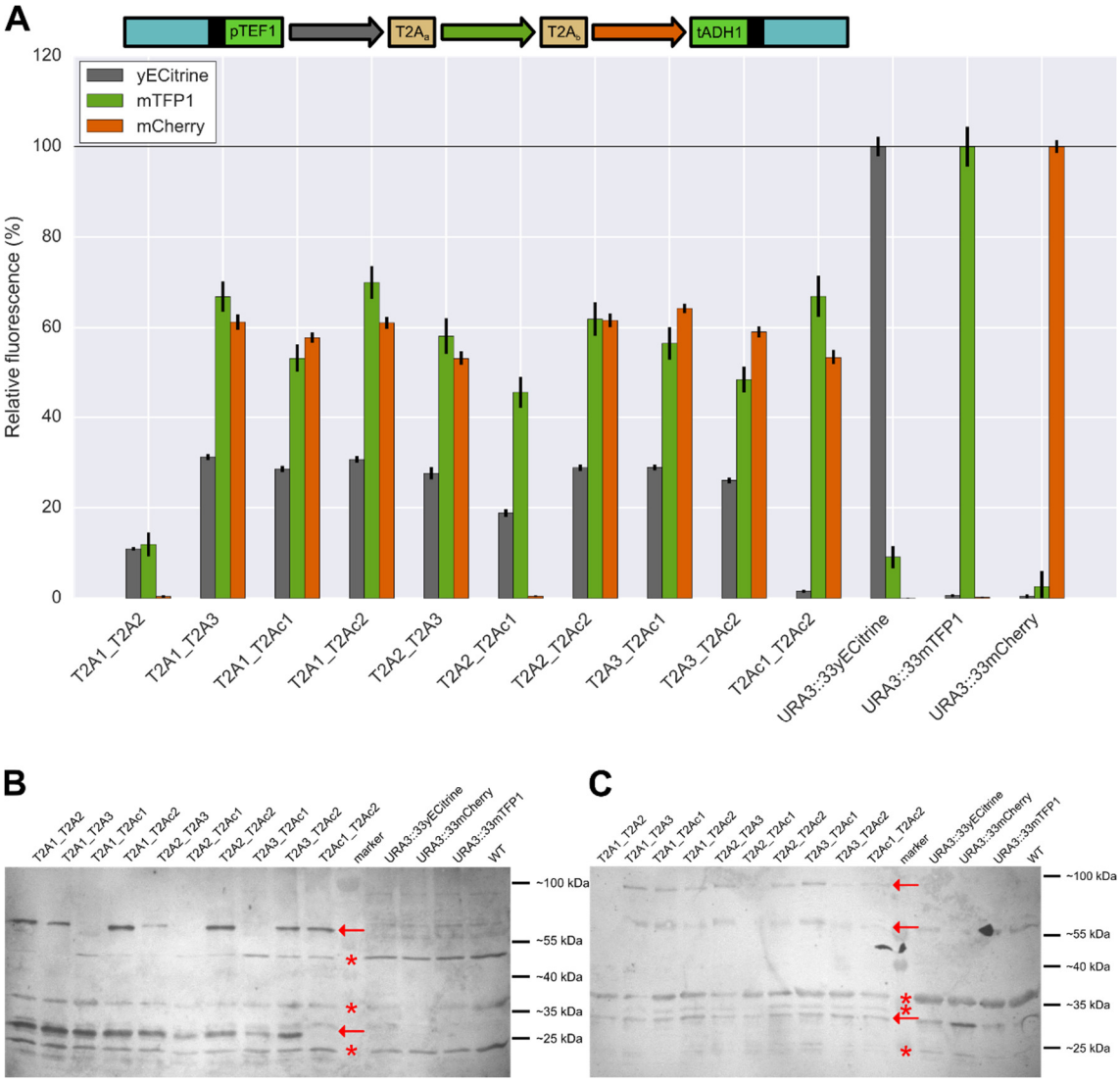

Figure 5.5: Tricistronic constructs integrated at the *URA3* locus using all combinations of designed T2A peptides. The strains corresponding with the represented combinations of T2A peptides are listed in Supplementary Table S.3.1. (A) Fluorescence of yECitrine, mTFP1 and mCherry as a measure for gene expression was normalized to their monocistronic reference strains (represented by the horizontal line). Error bars represent the standard error of the mean of three biological replicates. (B) Western blot for 2A-tagged protein detection using anti-2A. (C) Western blot for mCherry detection using anti-mCherry. Red arrows represent cleaved (bottom) and uncleaved (top, middle) protein products respectively, asterisks indicate unknown detected byproducts.

As such, polycistronic expression with an increasing number of ORFs will lead to a gradual decline in protein synthesis explaining the reduction in fluorescence compared to the monocistronic expressing reference strains. Also in enniatin B production strains [299], monocistronic expression of pathway genes led to higher titers compared to the tricistronic expression strains, confirming these observations.

Furthermore, irregularity in expression pattern for tricistronic TUs was observed for three combinations of T2A peptides, *i.e.* T2A1_T2A2 (sT2A19), T2A2_T2Ac1 (sT2A24) and T2Ac1_T2Ac2 (sT2A28) (Figure 5.5A). The results of this experiment were confirmed in an independent replication where fluorescence was measured after inoculation with new single colonies (Supplementary Figure S.3.7). In the quadcistronic expression units mTFP1 fluorescence at position 3 showed strong variation: for the first fluorescence experiment no expression was observed while for the independent control, mTFP1 fluorescence could be measured for some T2A combinations (Figure 5.6A, bottom). Overall, no mCherry fluorescence was detected here. These results suggest that with an increasing number of T2As in the polycistronic construct, complete termination at the downstream positioned T2As can occur more often, which was also hypothesized by Geier *et al.* [286] and is in line with the high impact on translation of ribosome drop-off for longer mRNAs [304].

Western blots of the tricistronic transcripts to evaluate T2A splicing activity were difficult to interpret. While the different protein products were visible with anti-2A and anti-mCherry antibodies (Figure 5.5B and C, arrows), many bands of additional byproducts were observed (Figure 5.5B and C, asterisks). For the anti-2A Western blot (Figure 5.5B), bands corresponding to the single proteins (yECit-T2A$_a$ and mTFP1-T2A$_b$, 28.6 kDa) and to the yECit-T2A$_a$-mTFP1-T2A$_b$ fusion product (57.2 kDa) were visible suggesting splicing activity of the T2A peptides. However, even if only one T2A peptide demonstrates cleavage activity, a band at 28.6 kDa will be visible, which makes it hard to conclude if either both or just one of the two T2A peptides has better activity. On the other hand, the anti-mCherry Western blot (Figure 5.5C) also showed the total fusion product yECit-T2A$_a$-mTFP-T2A$_b$-mCherry (83.5 kDa) which means totally unspliced protein products were present. As this fusion protein was not visible in the anti-2A immunoblot, these data suggest the more difficult binding of anti-2A to internal 'locked' 2A peptides. This is also clearly shown in Supplementary Figure S.3.4 where bands of the yECit-T2A$_a$-mCherry fusion proteins are barely observed while these of the spliced proteins, with thus freely accessible T2A tags, are

obviously present. In general, with this methodology it is very complicated to reveal the splicing efficiency of both T2A peptides in the tricistronic TU since always a mixture of different single and fusion products will be observed.

For quadcistronic expression on anti-2A Western blot and beside unspecific signals of byproducts (Figure 5.6B, asterisks), only bands equivalent to single reporter proteins were present (Figure 5.6B, arrows). These bands most probably correspond to only yECitrine-$T2A_a$ and mTagBFP2-$T2A_b$, since, on the one hand, no bands were observed for strain sT2A35 (T2A2_T2A3_T2Ac1), which lacks yECitrine and mTagBFP2 fluorescence (Figure 5.6A, top), and, on the other hand, only yECitrine and mTagBFP2 activity is observed. It is however remarkable that no bands corresponding to yECit-$T2A_a$-mTagBFP2-$T2A_b$ (85.3 kDa) were seen as never 100% splicing was detected in bicistronic expression (Figure 5.2B, C and Figure 5.4B, C) and yECit-$T2A_a$-mTFP1-$T2A_b$ fusion products were observed for tricistronic expression (Figure 5.5B). This especially indicates some unreliability of Western blotting as method to assess splicing efficiencies of T2A peptides in quadcistronic, and probably longer multicistronic TUs. The fact that no mCherry protein was detected (Figure 5.6C) seemed logical as also no fluorescence was observed (Figure 5.6A).

Nevertheless, these data and more in particular the fluorescence measurements suggest that expression of tricistronic TUs on the genome is feasible in *S. cerevisiae*, which is consistent with earlier plasmid based expression studies in *S. cerevisiae* and *P. pastoris*, and a genomic based study in *A. niger* [286,288,299]. However, gene expression levels significantly drop compared to monocistronic transcripts and expression of the third positioned ORF is not always reliable. Quadcistronic expression further leads to lower expression levels of the first two proteins and huge variation in the expression of the third protein. Overall, both methods, *i.e.* fluorescence measurements and Western blot, could not give an explicit interpretation of effective splicing efficiencies of the different T2A peptides on the different positions in the tri – and quadcistronic TUs. As such, it is still unclear to which extent the remaining protein activity (indicated by fluorescence) is due to spliced and/or unspliced protein products. To this end, these experiments revealed info about the total reporter activity, however it could not be said if cleaved and uncleaved proteins contributed equally or totally different, *i.e.* respectively more and less or vice versa, to the total fluorescence levels. To solve this bottleneck, the usage of fluorescence microscopy enabling the detection and localization of single fluorescent reporters could be a great help [305,306].

In conclusion, while acceptable protein activity was observed for the tricistronic TUs, the expression of the fluorescent reporters on the genome of *S. cerevisiae* for the quadcistronic constructs was not in accordance with our expectations. Particularly with the view of T2A peptide usage in heterologous pathways for the production of *e.g.* secondary metabolites, having many genes and ORFs that are on average much bigger than fluorescent reporters, and given their huge complexity, the effectiveness of this approach in microbial cell factory engineering of *S. cerevisiae* is rather low. It could thus be more interesting to use preferably bicistronic or tricistronic constructs (with robust performing T2As) under control of for example a bidirectional promoter [307,308] for the construction of large biosynthetic pathways using 2A peptides, rather than long quad – or multicistronic TUs. Nevertheless, it was shown that multicistronic expression is possible in *S. cerevisiae* and as such can lead to a reduction of promoters and terminators needed in a pathway. For instance, by using bi – or tricistronic TUs, the number of promoters and terminators can be decreased by half or two-thirds, respectively. Especially for long pathways, polycistronic expression can be seen as an interesting alternative.

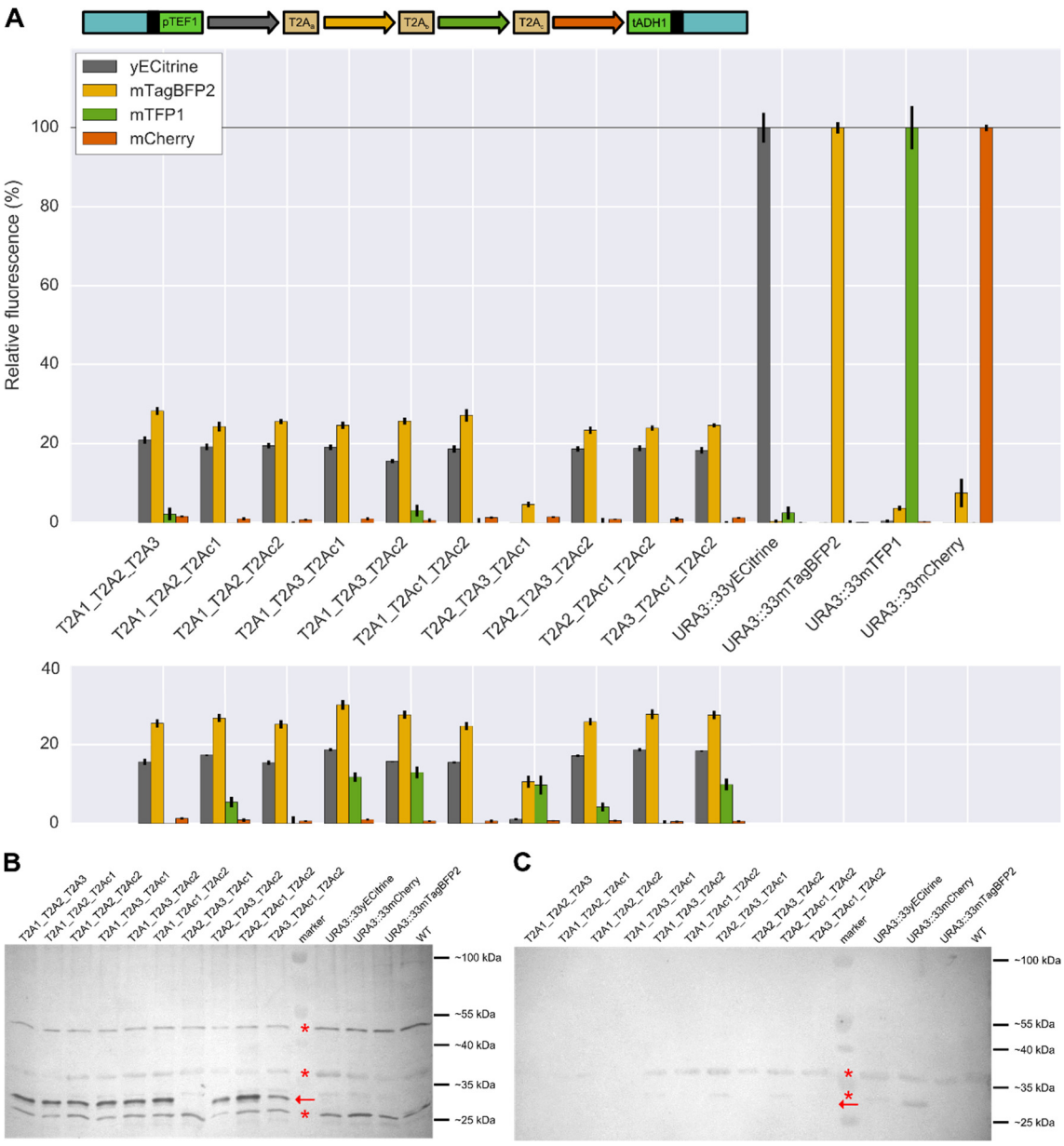

Figure 5.6: Quadcistronic construct integrated at the *URA3* locus using all combinations of designed T2A peptides. The strains corresponding with the represented combinations of T2A peptides are listed in Supplementary Table S.3.1. (A) Fluorescence of yECitrine, mTagBFP2, mTFP1 and mCherry as a measure for gene expression was normalized to their monocistronic reference strains (represented by the horizontal line). Also an independent fluorescence control experiment was performed (bottom). Error bars represent the standard error of the mean of three biological replicates. (B) Western blot for 2A-tagged protein detection using anti-2A. (C) Western blot for mCherry detection using anti-mCherry. Red arrows represent cleaved protein products, asterisks indicate unknown detected byproducts.

5.5 CONCLUSION

In this study, we thoroughly evaluated multicistronic gene expression in *S. cerevisiae* and report, for the first time, results for constructs integrated in the genome (*URA3* locus). Therefore, five 2A peptide sequences were designed based on the 2A sequence of the *Thosea asigna* virus. Their nucleotide composition differed as much as possible to avoid homologous recombination and ensure strain stability. Results of characterization experiments in bicistronic constructs revealed that these T2A peptides show cleavage activity in *S. cerevisiae* and can be used for multicistronic gene expression. One T2A peptide, *i.e.* T2Ac1, showed some lower and more unreliable splicing activity than the others. Even though, the palette of T2A peptides available for *S. cerevisiae* was successfully extended.

Next, double or triple combinations of the T2A peptides were used for the construction of respectively tri – or quadcistronic transcription units expressing fluorescent reporters. Stable integration in the genome was achieved since, based on colony PCR results, no homologous recombination between different T2A sequences was observed. Though in the tricistronic constructs relatively high fluorescence was obtained, protein activity was significantly lower compared to monocistronic expression and further decreased with increasing transcript length. These observations are in line with earlier studies that found a notable effect of ribosome drop-off in long mRNAs [304] and a negative correlation between mRNA length and its stability [302,303]. However, the used methods were insufficient to conclude anything about splicing efficiencies of the different positioned T2A peptides and to indicate the contribution of spliced and unspliced proteins to the total fluorescent activity in tri – and quadcistronic expression units. In this view, fluorescence microscopy could be a valuable technique for further research in this field.

In general, polycistronic expression of biosynthetic pathways on the genome of *S. cerevisiae* is achievable with our designed T2A peptides, but it remains unclear to which extent the total observed expression is caused by cleaved and uncleaved proteins. Additionally, the number of CDSs in one transcription unit must preferably be limited to two or three under control of for instance a bidirectional promoter, especially when large and complex heterologous pathway genes are used and sufficient amounts of enzymes are needed for efficient production. Nevertheless, multicistronic expression was proven to be a workable alternative to decrease the promoter and terminator usage in long pathways. In future work, it would therefore be interesting to verify if our T2A based pathway approach is

functional for the production of economically relevant molecules such as secondary metabolites like terpenoids and flavonoids.

CHAPTER 6 METABOLIC ENGINEERING OF *SACCHAROMYCES CEREVISIAE* INTO A PLATFORM STRAIN FOR THE PRODUCTION OF FLAVONOIDS

6.1 ABSTRACT 117
6.2 INTRODUCTION 118
6.3 MATERIAL AND METHODS 124
6.3.1 Strains and media 124
6.3.2 Construction of expression vectors for flavonoid biosynthesis 124
6.3.3 Plasmid construction for gene knock-outs and CRISPR/Cas9 127
6.3.4 Strain construction 128
6.3.5 Cultivation of yeast production strains 131
6.3.6 Detection and quantification of flavonoids and intermediates 131
6.3.7 Data analysis 132
6.4 RESULTS AND DISCUSSION 133
6.4.1 Design of the flavonoid pathway 133
6.4.2 Evaluating p-coumaric acid production in S. cerevisiae 133
6.4.2.1 Batch versus fed-batch conditions for p-coumaric acid production 135
6.4.2.2 Tyr route versus Phe-Tyr route for p-coumaric acid production 136
6.4.2.3 Effect of an enhanced flow to aromatic amino acids on p-coumaric acid production 136
6.4.3 Effect of an enhanced malonyl-CoA pool on naringenin production 140
6.4.4 De novo production of naringenin: combining enhanced flows toward aromatic amino acids and malonyl-CoA 142
6.5 CONCLUSION 147

Authors:

Thomas Decoene, Yatti De Nijs, Tom Delmulle, Nathalie Cuypers, Sofie De Maeseneire and Marjan De Mey

This chapter has been submitted as:

Decoene, T., De Nijs Y., Delmulle, T., Cuypers, N., De Maeseneire, S. L., and De Mey, M. (2018). Engineering the native precursor pools of Saccharomyces cerevisiae for the sustainable production of flavonoids: a naringenin case study. *Biotechnology and Bioengineering*.

Author contributions:

TDC, SDM and MDM were involved in the conception and design. TDC, SDM and MDM drafted the manuscript. TDC, YDN, TDM and NC were involved in the pathway and strain construction. Growth experiments, data analysis and interpretation of the results were performed by TDC.

6.1 ABSTRACT

Flavonoids are secondary metabolites naturally produced by plants with a lot of interesting biological properties making them applicable in the pharmaceutical and agricultural industry. Since their extraction from plants and chemical synthesis is inefficient and non-sustainable, microbial production is considered a worthy alternative to deliver these molecules.

In this study, we engineered *Saccharomyces cerevisiae* for the *de novo* biosynthesis of naringenin from glucose and evaluated its production capacity in two different culture conditions. In a first step, the phenylalanine and tyrosine precursor pools were engineered by alleviating negative feedback mechanisms and by deleting competing by-product formation, leading to enhanced p-coumaric acid titers (max. 161.91 ± 4.90 mg/l). Next, the two strain backgrounds with the highest p-coumaric acid production were selected to evaluate the effect of engineered cytosolic malonyl-CoA precursor supply on naringenin production. Therefore, an acetyl-CoA carboxylase (ScACC1p) with deregulated posttranslational phosphorylation was overexpressed, resulting in a final titer of 12.96 ± 0.62 mg/l naringenin when fed with p-coumaric acid. Finally, both approaches were combined for the *de novo* production of naringenin in a yeast strain optimized in its three flavonoid precursor pools. The highest naringenin titer we obtained was 4.07 ± 0.24 mg/l on deepwell MTP scale. Our strategy led to a 1.7 and 7.0-fold improvement in naringenin production compared to the non-optimized flavonoid precursor strain in batch and fed-batch conditions, respectively.

In conclusion, optimizing the flavonoid precursor pools in *Saccharomyces cerevisiae* is an attractive way to enhance naringenin production. As no pathway balancing was performed yet, optimization of the naringenin pathway itself is needed to further enhance production titers. In this view, our developed strain is a valuable chassis for the further development of yeast cell factories for flavonoid production.

6.2 INTRODUCTION

Amongst the large group of secondary plant metabolites, flavonoids gain more and more attention as target molecules in biological research since they have several interesting biological properties, including antioxidant, antibacterial, anti-inflammatory, antiviral and anticancer activities [44]. As such, flavonoids are an important group of compounds for the pharmaceutical and agricultural industry. Flavonoids naturally occur in plants and are synthesized via the phenylpropanoid pathway [43] which is essential for the production of monolignols, the building blocks of lignin [309]. These compounds belong to the large group of phenylpropanoids and based on their chemical structure, flavonoids are categorized into three main classes, *i.e.* bioflavonoids (2-phenylbenzopyrans), isoflavonoids (3-phenylbenzopyrans) and neoflavonoids (4-phenylbenzopyrans) (Figure 6.1) [310]. As natural producers, plants could be seen as a valuable source of these specialty metabolites, yet todays extraction methods are inefficient and often lead to mixtures of different phenylpropanoid compounds. In addition, chemical synthesis suffers from harsh reaction conditions and the difficulty of chiral centers. Therefore, microbial biosynthesis of flavonoids could be a compelling alternative to provide these compounds in sufficient amounts.

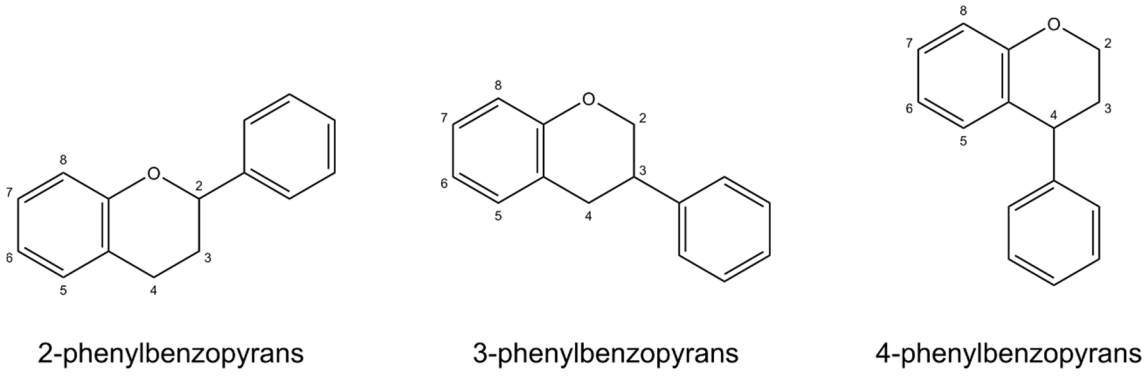

Figure 6.1: The three main classes of flavonoid structures classified according to their chemical structure [310].

Typically, the phenylpropanoid pathway in plants starts with the conversion of phenylalanine to cinnamic acid, and further to p-coumaric acid (Phe route). Subsequently, p-coumaric acid is activated by the addition of a coenzyme A group by 4-coumarate-CoA ligase, and converted into p-coumaroyl-CoA [311]. In flavonoid biosynthesis, p-coumaroyl-CoA

is modified into a chalcone (*e.g.* naringenin chalcon) by condensation with three activated malonyl-CoA molecules, catalyzed by a chalcone synthase. From such chalcone scaffolds, all possible flavonoids can be synthesized (Figure 6.2, blue part). As the phenylpropanoid pathway is not naturally available in microbes, plant-derived enzymes must be integrated in their native metabolism. Initial problems, especially in bacteria, with the need for a cytochrome P450-dependent cinnamate-4-hydroxylase (present in the phenylalanine dependent flavonoid pathway in plants), could be circumvented using bacterial tyrosine ammonia lyases [212,241], directly transforming tyrosine into p-coumaric acid through deamination. Thus, the tyrosine dependent pathway (Tyr route) can be seen as an extra route toward flavonoid production as p-coumaric acid can now be formed from both aromatic amino acids phenylalanine and tyrosine. In this regard, yeast is well suited for the heterologous production of flavonoids as it can functionally express both the bacterial and the plant-derived enzymes, and hence both pools can be harvested for flavonoid production.

For the efficient production of flavonoids in *S. cerevisiae*, an optimized pool turnover of precursor molecules is required, *i.e.* p-coumaric acid derived from phenylalanine and/or tyrosine, and malonyl-CoA (Figure 6.2, black part). Together with tryptophan, phenylalanine and tyrosine are synthesized via the shikimate pathway to serve as building blocks for protein synthesis [312]. The pathway is tightly regulated and starts with the condensation of phosphoenol pyruvate (PEP) and erythrose-4-phosphate (E4P) to 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP). PEP and E4P are derived from the glycolysis and pentose phosphate pathway (PPP), respectively. The condensation is catalyzed by the DAHP synthases Aro3p and Aro4p, which are negatively feedback regulated by respectively phenylalanine and tyrosine [48,313]. Further on, the pentafunctional enzyme Aro1p catalyzes the five central reactions in the shikimate pathway toward 5-enolpyruvylshikimate-3-phosphate (EPSP), which is further converted to chorismate by a chorismate synthase encoded by *ARO2*. From here, the pathway is split into the tryptophan and the phenylalanine-tyrosine branches. In the latter, the last common intermediate for the important flavonoid precursors phenylalanine and tyrosine, *i.e.* prephenate, is formed through a Claisen rearrangement by the Aro7p chorismate mutase. Also this enzyme is strongly feedback regulated, as it is inhibited by tyrosine and activated by tryptophan. Finally, prephenate leads either to phenylalanine through prephenate dehydratase Pha2p and aromatic aminotransferases I and II (Aro8p and Aro9p) with the phenylpyruvate intermediate, or to tyrosine through a prephenate dehydrogenase Tyr1p forming p-

hydroxyphenylpyruvate which again is further modified by the aminotransferases encoded by *ARO8* and *ARO9*.

Since the DAHP synthases (Aro3p and Aro4p) and the chorismate mutase (Aro7p) are negatively influenced by the flavonoid precursors phenylalanine and tyrosine, enzyme engineering could contribute to a continuous flux to these precursors for the sustainable production of flavonoids. Indeed, it was demonstrated that eliminating the negative feedback mechanisms of DAHP synthase and chorismate mutase improved the flux toward phenylalanine and tyrosine and hence increased the production of their downstream derivatives [5,48,55,64,65,313]. More specifically, DAHP synthase activity is modified by deleting the *ARO3* gene and creating a tyrosine insensitive ARO4p (Aro4p^{G226S}) [48,224], and precursor production is further improved using an aromatic amino acid insensitive chorismate mutase (Aro7p^{G141S}) [64–66,314]. Also, the deletion of pyruvate decarboxylases responsible for the production of aromatic alcohols from phenylpyruvate (Pdc5p, Pdc6p, Aro10p) improved the pathway flux to phenylalanine [48,65].

Malonyl-CoA, another important precursor for flavonoid production, is an essential precursor for the biosynthesis of fatty acids and thus its pathway is also under strong metabolic control. Cytosolic malonyl-CoA is formed out of acetyl-CoA by acetyl-CoA carboxylase, encoded by *ACC1*. Its transcription is regulated positively and negatively by the transcription factors Ino2p/4p and Opi1p, respectively. Additionally, posttranslational phosphorylation occurs by Snf1p, which decreases Acc1p activity under decreased acetyl-CoA levels [78,315,316]. As Snf1p also plays an important role in the regulation of other cellular processes, removing the *SNF1* gene is not a sensible option. On the other hand, it was reported that phosphorylation of Acc1p can be avoided when the putative phosphorylation sites at Ser659 and Ser1157 are changed to an alanine [78]. Such mutations obviously led to a higher production of malonyl-CoA-derived compounds explained by the higher Acc1p activity [5,78,267]. In addition, overexpression of the Acc1p enzyme also proved to enhance the cytosolic malonyl-CoA pool in *S. cerevisiae* [77]. Finally, the biosynthesis of malonyl-CoA can also be improved by optimizing the intracellular pool of acetyl-CoA [55,316,317].

The aforementioned metabolic engineering strategies, extensively reviewed by Delmulle *et al.* [13], led to the improved production of several phenylpropanoid compounds or their intermediates. To date, mainly strains with improved aromatic amino acid pools yielding increased titers of p-coumaric acid [65,297] and flavonoids [48,49] are evaluated. For example,

tyrosine derived p-coumaric acid titers up to 1.93 g/l were obtained by optimizing the flow toward aromatic amino acids [65]. Similarly, improved naringenin titers ranging from 1.55 mg/l [49] to 109 mg/l [48] were obtained using the Tyr route versus the Phe-Tyr route, respectively, in strains with optimized aromatic amino acid pools. By our knowledge, only in two studies the production of resveratrol is evaluated by combining an improved p-coumaric acid pool through an improved Tyr route or an improved Phe route with an improved malonyl-CoA pool, which led to resveratrol titers of 235.57 mg/l [5] and 272.64 mg/l [267] respectively. Up to date, no reports have been published regarding the biosynthesis of flavonoids that combine strategies to enhance the flow to aromatic amino acid pools via both the Phe and the Tyr route (Phe-Tyr route) with the enhanced production of cytosolic malonyl-CoA. In this study, we therefore evaluated different yeast strain backgrounds optimized for one, two or all three of the main flavonoid precursors, phenylalanine, tyrosine or malonyl-CoA, for the production of p-coumaric acid and naringenin, under different culture conditions. The impact of eliminating feedback inhibition and deleting by-product formation on the biosynthesis of p-coumaric acid using the Tyr or the Phe-Tyr route was investigated first. Next, the influence of an enhanced malonyl-CoA pool on the production of naringenin was examined in cultures fed with p-coumaric acid. Finally, improved malonyl-CoA precursor production was engineered in the best p-coumaric acid producers to assess their *de novo* naringenin production. To the best of our knowledge, this is the first study that employs both the plant and bacterial pathway for p-coumaric acid production in combination with an enhanced cytosolic malonyl-CoA pool to assess the effect on the production of a flavonoid, both in batch and fed-batch conditions.

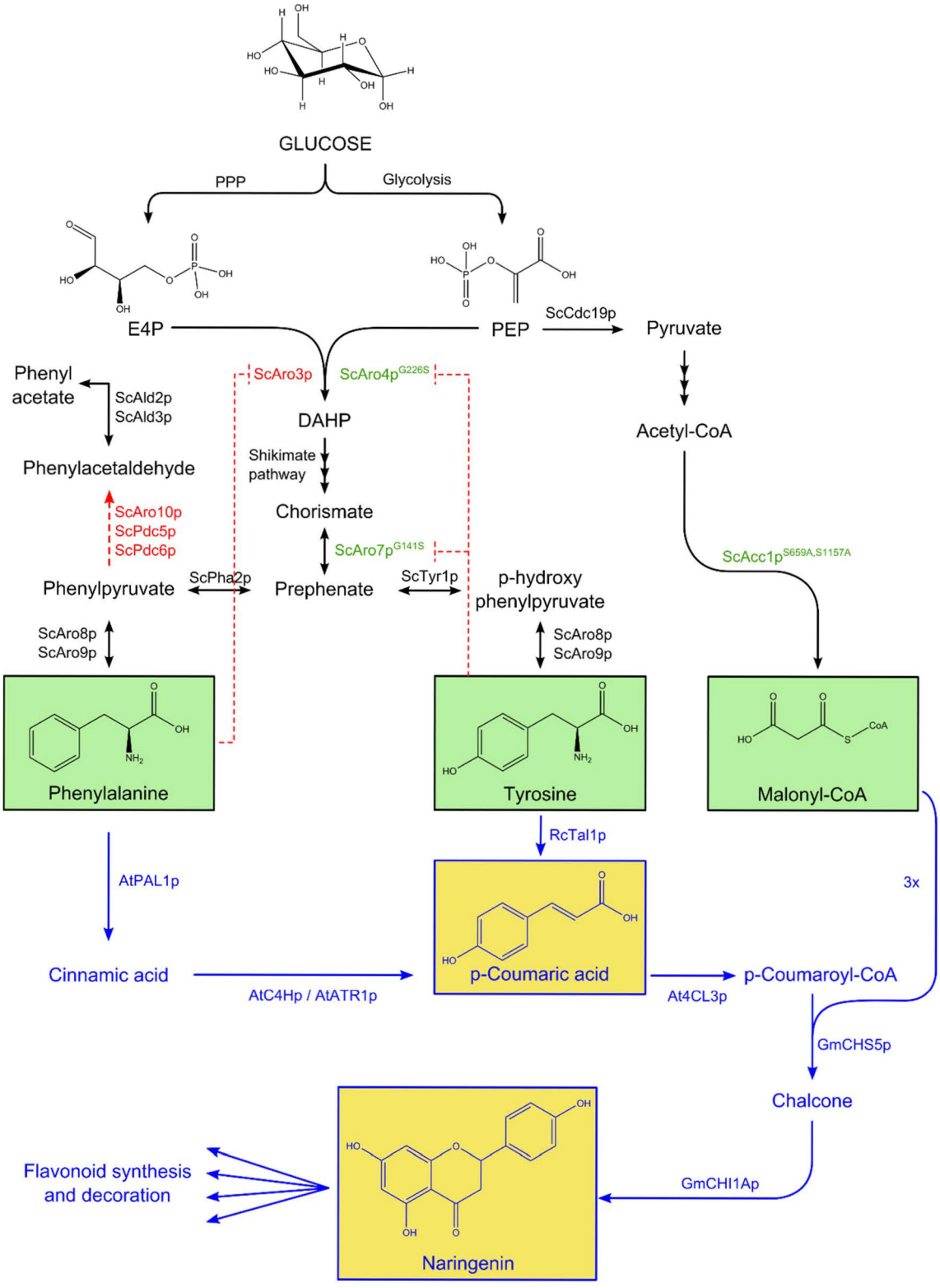

Figure 6.2: Schematic representation of the *de novo* flavonoid biosynthesis pathway in *S. cerevisiae* implemented for this study. Native enzymes and intermediates are indicated in black, gene knockouts are indicated in red and enzymes mutated to enhance the flavonoid precursor pools are shown in green. The red dashed lines represent the negative feedback inhibition of

phenylalanine and tyrosine, and the loss in phenylalanine through by-product formation toward phenylacetaldehyde. This is circumvented by the indicated gene knock-outs and mutated enzymes. The heterologous flavonoid pathway is highlighted in blue. Optimized flavonoid precursor pools are indicated in a green box, the intermediate p-coumaric acid and the end product naringenin are indicated in a yellow box. PPP: Pentose Phosphate Pathway; E4P: erythrose-4-phosphate; PEP: Phosphoenol pyruvate; DAHP: 3-deoxy-D-arabino-heptulosonate-7-phosphate; ScCdc19p: Pyruvate kinase; ScAro3p/ScAro4p: DAHP synthase; ScAro7p: chorismate mutase; ScTyr1p: prephenate dehydrogenase; ScPha2p: prephenate dehydratase; ScAro8p/ScAro9p: aromatic aminotransferases I/II; ScAro10p, ScPdc5p, ScPdc6p: pyruvate decarboxylases; ScAcc1p: acetyl-CoA carboxylase; ScAld2p/ScAld3p: aldehyde dehydrogenases; AtPAL1p: phenylalanine ammonia-lyase; RcTal1p: tyrosine ammonia-lyase; AtC4Hp: cinnamate-4-hydroxylase; AtATR1p: cytochrome-P450-reductase; At4CL3p: 4-coumarate-CoA ligase; GmCHS5p: chalcone synthase; GmCHI1Ap: chalcone isomerase. Enzyme prefixes: At: *Arabidopsis thaliana*; Gm: *Glycine max*; Rc: *Rhodobacter capsulatus*; Sc: *Saccharomyces cerevisiae*.

6.3 MATERIAL AND METHODS

Unless otherwise stated, all products were purchased from Sigma-Aldrich (Diegem, Belgium), CPEC [11] was used for the assembly of plasmids and plasmid extraction was performed with the innuPREP Plasmid Mini Kit (Analytik Jena AG, Jena, Germany).

6.3.1 Strains and media

Transformax™ EC100™ Electrocompetent *E. coli* (Lucigen, Halle-Zoersel, Belgium) or DH5α™ *E. coli* (ThermoFisher Scientific, Aalst, Belgium) was used for cloning procedures and for maintaining plasmids. *E. coli* strains were cultured in lysogeny broth (LB) consisting of 1% tryptone-peptone (Difco, Erembodegem, Belgium), 0.5% yeast extract (Difco), 0.5% sodium chloride (VWR, Leuven, Belgium) and 100 µg/ml ampicillin or 25 µg/ml chloramphenicol dependent on the selection marker. For the selection of *E. coli* strains after Golden Gate, sucrose medium without salt existing of 1% tryptone-peptone (Difco), 0.5% yeast extract (Difco) and 5% sucrose was used. For solid growth medium, 1% agar (Biokar diagnostics, Pantin Cedex, France) was added.

S. cerevisiae SY992 (*Matα, ura3Δ0, his3Δ1, leu2Δ0, trp1-63, ade2Δ0, lys2Δ0, ADE8*) (Euroscarf [205]) was used as host for flavonoid production. All yeast strains used in this study are derived from this strain. They are listed in Table 6.2. Yeast strains were maintained/selected on synthetic defined (SD) medium consisting of 0.67% YNB without amino acids, 2% glucose (Cargill, Sas van Gent, The Netherlands) and selective amino acid supplement mixture (MP Biomedicals, Brussel, Belgium) dependent on the required auxotrophies. For solid media, 2% Agar Noble (Difco) was added. SD medium as well as synthetic fed-batch medium were used to evaluate the different p-coumaric acid and naringenin production strains. Feed-In-Time (FIT) synthetic fed-batch medium M-Sc.syn-1000 was ordered from M2P labs (Baesweiler, Germany). Prior to use, an enzyme mix (final concentration of 0.5% v/v) and a vitamin mix (final concentration of 1% v/v) was added to the Sc.syn Base solution. If needed, extra p-coumaric acid was added to the medium with a final concentration of 164.05 mg/l (1mM).

6.3.2 Construction of expression vectors for flavonoid biosynthesis

An overview of all parts used to construct the expression vectors enabling flavonoid biosynthesis in *S. cerevisiae* SY992 is given in Table 6.1. Except for *AtATR1,* where we used

the native sequence, all genes for the flavonoid pathway were codon harmonized for *S. cerevisiae* with the EuGene software [318] (harmonization performed by RSCU, minimizing free energy of secondary RNA structures and avoiding Kozak sequence motifs and BsaI sites). The harmonized genes were ordered as gBlocks® from Integrated DNA Technologies (IDT, Leuven, Belgium) and are listed in Supplementary Table S.4.1. All promoters and native terminators were PCR-amplified from *S. cerevisiae* SY992 genomic DNA using PrimeSTAR HS DNA polymerase (Takara, Westburg, Leusden, The Netherlands). The synthetic terminators were ordered as DNA oligonucleotides from IDT.

Table 6.1: Parts used for the assembly of yeast expression plasmids for the flavonoid pathway. The sequences of the left (L_VA) and right (R_VA) VEGAS adapters were obtained from Kuijpers *et al.* (2013) [290].

L_VA	Promoter	CDS	Enzyme name	Uniprot	Terminator	R_VA
LVI	pTDH3 [192]	*AtPAL1*[a]	Phenylalanine ammonia lyase	P35510	tENO1 [192]	RVA
LVA	pPGK1 [192]	*AtC4H*	Cinnamate-4-hydroxylase	B1GV49	tSynth9 [24]	RVB
LVB	pSAC6 [192]	*AtATR1*	Cytochrome P450-reductase	Q9SB48	tGUO1 [24]	RVC
LVC	pTEF1 [159]	*RcTal1*[b]	Tyrosine ammonia lyase	NA –	tADH1 [206]	RVD
LVI	pTDH3 [192]	*At4CL3*	4-coumarate-CoA ligase	Q9S777	tGUO1 [24]	RVF
LVF	pPGK1 [192]	*GmCHS5*[c]	Chalcone synthase	P48406	tSynth17 [24]	RVG
LVG	pTIF6 [d]	*GmCHI1A*	Chalcone isomerase	Q93XE6	tSynth18 [24]	RVJ
L_VA	**Essential plasmid maintenance elements**					**R_VA**
LVH	CEN6/ARS4					RVI
LVD	AmpR-pMB1ori-pAgTEF1_*SpHIS5*_tAgTEF1 [319]					RVH
LVJ	AmpR-pMB1ori-pKlLEU2_*KlLEU2*_tAgTEF1 [319]					RVH

[a] At: *Arabidopsis thaliana;* [b] Rc: *Rhodobacter capsulatus;* [c] Gm: *Glycine max;* [d] intergenic non-coding region of the essential Translation Initiation Factor 6 (SGD 593069 to 593486, chrXVI); Ag: *Ashbya gossypii*; Kl: *Kluyveromyces lactis*. NA: Not Available.

The different transcription units (TUs) for the flavonoid biosynthesis pathway were assembled by yeast Golden Gate (yGG) [320]. Therefore, all parts (*i.e.* promoters, coding sequences, terminators and adapters) were flanked by inward-facing BsaI sites and were assembled to a yGG carrier vector carrying a BsaI insensitive ampicillin resistance gene. The yGG destination vector contained a chloramphenicol resistance marker and outward-facing BsaI sites flanking a *SacB* gene which is replaced by correctly assembled TUs and enables screening of correct *E. coli* colonies on sucrose medium without salt [207]. In addition, inward-facing AarI sites were introduced outside the BsaI sites of the destination vector for TU excision. Adapters were integrated in the yeast TUs to facilitate plasmid construction by *in vivo* recombination. Also a carrier vector with inward-facing AarI sites consisting of only a

CEN6/ARS4, and different carrier vectors with a particular yeast auxotrophic marker, a pMB1 ori and a BsaI insensitive ampicillin resistance marker were constructed (Table 6.1). These carrier vectors were used as sources of the elements for replication and selection of yeast vectors in *E. coli* and *S. cerevisiae*.

For the Golden Gate assembly of TUs in the yGG destination vector, 100 ng yGG destination vector with equimolar amounts of every part-containing carrier vector were mixed in a one-pot Golden Gate reaction. The one-pot restriction-ligation reaction was performed as described by Agmon *et al.* [320] but 20U of BsaI (NEB, R3535L) and 400U of T4 DNA ligase (NEB, M0202L) were used. 5 µl of yGG reaction mixture was chemically transformed in DH5α™ *E. coli* cells, which were plated on salt-lacking sucrose plates containing 25 µg/ml chloramphenicol. Growing colonies were confirmed by colony PCR and plasmids were verified by sequencing (EZ-Seq, Macrogen, Amsterdam, The Netherlands).

The final flavonoid expression vectors pCoumPT and pNar (Supplementary Table S.4.2) were assembled via PCR-mediated VEGAS [21]. TUs were PCR-amplified from their respective yGG destination plasmids and essential plasmid elements from their carrier vectors by 20 bp primers annealing at the ends of the left and right adapters. These 60 bp adapters at each side of the TU served as homologous overlap for *in vivo* recombination in yeast. Yeast transformations were carried out with the lithium-acetate method [208]. 200 fmol of every TU and 100 fmol of each plasmid maintenance element were used in a total volume of 34 µl. After transformation, cells were selected on SD medium lacking histidine or leucine for 3-4 days at 30°C. Correct overlaps were confirmed by yeast colony PCR and plasmids were extracted with a user developed protocol from the QIAprep Spin Miniprep Kit (*'Isolation of plasmid DNA from yeast using the QIAprep Spin Miniprep Kit'*, QIAGEN, Antwerp, Belgium). The *in vivo* assembled flavonoid expression vectors were further transformed in Transformax™ EC100™ Electrocompetent *E. coli* (Lucigen) and confirmed by sequencing (EZ-Seq, Macrogen).

The expression vector for p-coumaric acid production only using the *RcTal1* gene (pCoumT) was constructed by using the TU amplified from the *RcTal1* destination vector and an in-house low-copy *URA3* backbone. To make the Acc1p^{S659A,S1157A} overexpression vector (pOEACC1^{S659A,S1157A}), the native *ACC1* coding sequence was first picked up from genomic DNA of *S. cerevisiae* SY992 by a PrimeSTAR HS DNA polymerase PCR (Takara) and assembled in an in-house low-copy *URA3* yeast backbone. Subsequently, mutations to

replace the amino acid codons from serine to alanine at positions 659 and 1157 in Acc1p were introduced by PCR using mismatching primers (IDT), followed by CPEC to assemble the final Acc1p^{S659A,S1157A} overexpression vector. Expression vectors for the negative control strains (pURA3, pHIS5 and pLEU2) only contained the auxotrophic marker TU and the essential plasmid maintenance elements. After plasmid confirmation via sequencing (EZ-Seq, Macrogen), all expression vectors were transformed in the appropriate yeast strains using the lithium-acetate method [208].

6.3.3 Plasmid construction for gene knock-outs and CRISPR/Cas9

Knock-out cassettes were constructed by flanking the auxotrophic marker genes of the pBN100, pUG27 and pUG73 deletion marker plasmids from Euroscarf [319,321,322] with the respective 500 bp up – and downstream homologies from the coding sequence of the knock-out of interest. The up – and downstream homologies were PCR-amplified from *S. cerevisiae* SY992 genomic DNA. All three pieces were assembled into a pJET backbone (ThermoFisher Scientific).

The gRNA expression plasmids needed for genomic alterations in *ARO4* and *ARO7* using CRISPR/Cas9 were constructed from p426-SNR52p-gRNA.CAN1.Y-SUP4t (Addgene #43803) [19]. For easy selection of right clones after CPEC in *E. coli*, a template plasmid was made, p426-SNR52p-*aeBlue*-SUP4t, where *aeBlue* (iGEM part BBa_K864401) replaces the original gRNA sequence. This vector was then used as template to amplify the gRNA expression backbone by PCR. As such, white colonies were obtained after correct integration of the gRNA in the gRNA expression backbone. Used gRNA sequences for the mutations in Aro4p^{G226S} and Aro7p^{G141S} were respectively 5' tgctcattctcaccatttca 3' and 5' ggtgatgataagaataactt 3', and were selected using the CRISPy tool (http://staff.biosustain.dtu.dk/laeb/crispy_cenpk/) [323]. These gRNAs were ordered as 60bp oligonucleotides (IDT) where the 20bp gRNA sequence was flanked at each side with 20bp compatible backbone ends for CPEC. For the construction of CRISPR/Cas9 donor DNA template plasmids, a similar approach was used as for the Acc1p^{S659A,S1157A} overexpression vector. First, the native *ARO4* and *ARO7* coding sequences were picked up from genomic DNA of *S. cerevisiae* SY992 and assembled in an ampicillin resistant *E. coli* plasmid backbone. Subsequently, mutations to replace the amino acid codons from glycine to serine at positions 226 and 141 in Aro4p and Aro7p respectively were introduced by PCR using

mismatching primers (IDT), followed by CPEC to assemble the donor DNA template plasmid and sequencing (EZ-Seq, Macrogen) for plasmid verification.

All plasmids used in this study are listed in Supplementary Table S.4.2 and an overview of primers used for amino acid modifications is listed in Supplementary Table S.4.3.

6.3.4 Strain construction

For the construction of strains with gene knock-outs, knock-out cassettes were PCR-amplified from their respective template plasmids (Supplementary Table S.4.2) and transformed as linear DNA in the appropriate *S. cerevisiae* strains according to the lithium-acetate method [208]. Afterwards, the *HIS5* and *LEU2* auxotrophic markers were removed by the *Cre-loxP* recombination system using pSH47 as earlier described [319] and *URA3* markers were eliminated by selection on SD medium containing 0.1% 5-fluoroorotic acid (FOA). The correct genomic integration of knock-out cassettes and removal of auxotrophic markers was verified by yeast colony PCR.

For the introduction of genomic mutations in *ARO4* and *ARO7*, the Cas9 expression vector p414-TEF1p-Cas9-CYC1t (Addgene #43802) was transformed in the appropriate strains by the lithium-acetate method [208]. Next, 1 µg of gRNA plasmid (p426ARO4 or p426ARO7) with 1 pmol of linear PCR-amplified donor DNA for introducing the $ARO4^{G226S}$ or $ARO7^{G141S}$ mutations was transformed via the lithium-acetate method [208]. For the simultaneous insertion of both $ARO4^{G226S}$ and $ARO7^{G141S}$ mutations, the earlier reported CRISPR/Cas9 method with linearized gRNA plasmid backbone and linear gRNA cassettes was used (gap repair method) [298]. Therefore, 150 ng gRNA plasmid backbone with 400 ng of each linearized gRNA cassette and 600 ng of the proper donor DNA was transformed. After transformation, strains were selected on SD medium. The correct introduction of the $ARO4^{G226S}$ and $ARO7^{G141S}$ mutations was confirmed with sequencing (EZ-Seq, Macrogen). Afterwards, the Cas9 expression vector and gRNA plasmids were removed by growing the strains on non-selective SD medium according to the 'Plasmid Loss Assay' protocol (OpenWetWare).

The $ACC1^{S659A,S1157A}$ overexpression strains and flavonoid production strains were constructed by transformation of pOEACC1^{S659A,S1157A}, pCoumT, pCoumPT and/or pNar in the appropriate yeast strains [208]. Correct strains were verified by colony PCR. An overview of all strains constructed and used in this study is given in Table 6.2.

Table 6.2: *S. cerevisiae* strains used in this study.

Strain	Genotype	Plasmids	Reference
SY992	*Matα, ura3Δ0, his3Δ1, leu2Δ0, trp1-63, ade2Δ0, lys2Δ0, ADE8*	-	205
sCoumPT01	SY992	pCoumPT	This study
sCoumPT02	SY992 *ΔARO3*	pCoumPT	This study
sCoumPT03	SY992 *ΔARO10*	pCoumPT	This study
sCoumPT04	SY992 *ΔPDC5*	pCoumPT	This study
sCoumPT05	SY992 *ΔPDC6*	pCoumPT	This study
sCoumPT06	SY992 $ARO4^{G226S}$	pCoumPT	This study
sCoumPT07	SY992 $ARO7^{G141S}$	pCoumPT	This study
sCoumPT08	SY992 $ARO4^{G226S}$ $ARO7^{G141S}$	pCoumPT	This study
sCoumPT09	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10*	pCoumPT	This study
sCoumPT10	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10* $ARO4^{G226S}$	pCoumPT	This study
sCoumPT11	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10* $ARO7^{G141S}$	pCoumPT	This study
sCoumPT12	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10* $ARO4^{G226S}$ $ARO7^{G141S}$	pCoumPT	This study
sCoumPT13	SY992	pHIS5	This study
sCoumT01	SY992	pCoumT	This study
sCoumT02	SY992 *ΔARO3*	pCoumT	This study
sCoumT03	SY992 *ΔARO10*	pCoumT	This study
sCoumT04	SY992 *ΔPDC5*	pCoumT	This study
sCoumT05	SY992 *ΔPDC6*	pCoumT	This study
sCoumT06	SY992 $ARO4^{G226S}$	pCoumT	This study
sCoumT07	SY992 $ARO7^{G141S}$	pCoumT	This study
sCoumT08	SY992 $ARO4^{G226S}$ $ARO7^{G141S}$	pCoumT	This study
sCoumT09	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10*	pCoumT	This study
sCoumT10	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10* $ARO4^{G226S}$	pCoumT	This study
sCoumT11	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10* $ARO7^{G141S}$	pCoumT	This study
sCoumT12	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10* $ARO4^{G226S}$ $ARO7^{G141S}$	pCoumT	This study
sCoumT13	SY992	pURA3	This study

Strain	Genotype	Plasmids	Reference
sNar01	SY992	pCoumPT, pNar	This study
sNar02	SY992 *ARO4*^{G226S}	pCoumPT, pNar	This study
sNar03	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10 ARO4*^{G226S}	pCoumPT, pNar	This study
sNar04	SY992	pHIS5, pLEU2	This study
sNarA01	SY992	pCoumPT, pNar, pOEACC1^{S659A,S1157A}	This study
sNarA02	SY992 *ARO4*^{G226S}	pCoumPT, pNar, pOEACC1^{S659A,S1157A}	This study
sNarA03	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10 ARO4*^{G226S}	pCoumPT, pNar, pOEACC1^{S659A,S1157A}	This study
sNarA04	SY992	pHIS5, pLEU2, pURA3	This study
sNarC01	SY992	pNar	This study
sNarC02	SY992 *ARO4*^{G226S}	pNar	This study
sNarC03	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10 ARO4*^{G226S}	pNar	This study
sNarC04	SY992	pLEU2	This study
sNarAC01	SY992	pNar, pOEACC1^{S659A,S1157A}	This study
sNarAC02	SY992 *ARO4*^{G226S}	pNar, pOEACC1^{S659A,S1157A}	This study
sNarAC03	SY992 *ΔARO3 ΔPDC5 ΔPDC6 ΔARO10 ARO4*^{G226S}	pNar, pOEACC1^{S659A,S1157A}	This study
sNarAC04	SY992	pLEU2, pURA3	This study

6.3.5 Cultivation of yeast production strains

For the growth experiments with the p-coumaric acid and naringenin production strains, three biological replicates per strain were inoculated from agar plate in 200 µl selective SD medium in a sterile µclear, flat-bottomed, white 96-well microtiter plate (Greiner Bio-One, Vilvoorde, Belgium) enclosed by a Breathe-Easy® sealing membrane (Sigma-Aldrich). These pre-culture MTPs were grown for 24h on a Compact Digital Microplate Shaker (ThermoFisher Scientific) at 800 rpm and 30°C. For the main cultivation experiments MTPs with air-penetrable sandwich cover (EnzyScreen, Heemstede, The Netherlands) were used. 50 µl of the pre-culture was used for inoculating 500 µl medium in 96 deep-well MTPs (EnzyScreen) for the evaluation of p-coumaric acid and naringenin production fed with 164.05 mg/l (1mM) p-coumaric acid or 150 µl of the pre-culture was used for the inoculation of 3 ml medium in 24 deep-well MTPs (EnzyScreen) for *de novo* naringenin production. All cultivations were carried out for 72h at 30°C, and 350 rpm or 300 rpm (2.5 cm orbit) for 96 or 24 deep-well MTPs, respectively. At the end of cultivation, the optical density was measured at 600 nm (OD600) by diluting 15 µl culture in 135 µl deionized water in a µclear, flat-bottomed, black 96-well microtiter plate (Greiner Bio-One). The OD600 was determined in a TECAN Infinite® 200 PRO (Tecan) MTP reader. Afterwards, cultures were spun down and the supernatant was used for metabolite detection and quantification using Ultra Performance Liquid Chromatography (UPLC).

6.3.6 Detection and quantification of flavonoids and intermediates

Naringenin and intermediates such as p-coumaric acid, cinnamic acid and phloretic acid were measured using a Waters Acquity UPLC connected to a UV detector and equipped with a Kinetex® 2.6 µm Polar C18 column (Phenomenex, Utrecht, The Netherlands) operated at 30°C. A gradient method with two eluents, *i.e.* 13 mM trifluoroacetic acid (TFA) (A) and pure acetonitrile (ACN) (B), with a flow rate of 0.6 ml/min was used. The UPLC method started with 10% of eluent B, followed by a linear increase to 23% of eluent B (0 – 2.5 min) where its fraction was subsequently further increased to 70% (2.5 – 5.0 min). Next, the fraction was maintained at 70% of eluent B (5.0 – 6.0 min), finally the fraction of eluent B was decreased from 70% to 10% (6.0 – 8.0 min). Phloretic acid was detected at 277 nm and had a retention time of 1.9 min. p-coumaric acid, cinnamic acid and naringenin were detected at 290 nm with retention times of 2.3, 4.1 and 4.5 min, respectively. Peak areas were integrated with OpenChrom® and concentrations were determined from phloretic acid, p-

coumaric acid, cinnamic acid and naringenin standard curves. All standards were HPLC grade (> 95% purity) and purchased from Sigma-Aldrich.

6.3.7 Data analysis

Unless otherwise stated, all calculations were performed in Python using the Python Data Analysis Library (Pandas). Error bars represent the standard error of the mean (n = 3). Pairwise comparisons between different strains were done by a two-sided T-test using the scipy.stats package in Python. ANOVA was performed in SPSS Statistics 24, where normality was checked with the Shapiro-Wilk's Test and homoscedasticity with the Levene's Test (which was optional, as all populations tested had equal sample sizes). In all cases, a significance level of 0.05 was applied.

6.4 RESULTS AND DISCUSSION

6.4.1 Design of the flavonoid pathway

To optimize the *de novo* biosynthesis of flavonoids in *S. cerevisiae* starting from glucose, the pathway was split in an upstream and downstream part. The upstream part comprises the production of p-coumaric acid, either using only the tyrosine pool (*RcTal1*; Tyr route) or departing from both tyrosine and phenylalanine (*AtPAL1, AtC4H, AtATR1* and *RcTal1*; Phe-Tyr route). The downstream part of the pathway starts from p-coumaric acid leading to the end product naringenin (*At4CL3, GmCHS5* and *GmCHI1A*). This strategy allows to separately investigate the effects of the phenylalanine and tyrosine pools, and the malonyl-CoA pool on flavonoid biosynthesis in yeast. With the upstream module, it can be checked if the gene knock-outs and enzyme engineering strategies applied for an enhanced phenylalanine and tyrosine synthesis effectively lead to higher p-coumaric acid titers. In addition, a comparison can be made between using solely the tyrosine pool and using both precursor pools. With the downstream part of the pathway, and by feeding with p-coumaric acid, the impact of malonyl-CoA supply on naringenin production can be analyzed. Finally, both modules can be combined, for an optimized *de novo* naringenin production. In the current study, strong constitutive promoters and moderate strength terminators were used to express the flavonoid pathway genes, except for *GmCHI1A* where the medium strength pTIF6 promoter was chosen because of the high activity of GmCHI1Ap toward naringenin chalcone [324]. The modules were built on low copy expression vectors to minimize expression variability [20,192].

6.4.2 Evaluating p-coumaric acid production in S. cerevisiae

In literature, two main strategies are reported to enhance the flow toward aromatic amino acid pools: (1) enhancing the flux toward the common precursor prephenate by alleviation of the feedback inhibition on DAHP synthases (*e.g.* deletion of Aro3p and/or expression of Aro4p^{G226S}) and/or on chorismate mutase (*e.g.* expression of Aro7p^{G141S}) [5,48,49,65,297], and (2) decreasing the by-product formation of aromatic alcohols depleting the flux to aromatic amino acids by deleting (phenyl)-pyruvate decarboxylases (*e.g.* Aro10p, Pdc5p and Pdc6p) [48,49,65,297]. The influence of these alterations on p-coumaric acid and derived flavonoid production was mainly assessed using the bacterial pathway to p-coumaric acid (Tyr route). Here, these strategies were evaluated using either the bacterial pathway or both the plant

and bacterial pathway (Phe-Tyr route). To this end, *S. cerevisiae* strains were created carrying one of the above mentioned alterations or combinations thereof. These strains were transformed with the plasmid carrying the upstream part of the flavonoid production pathway to p-coumaric acid via the Tyr route (pCoumT, sCoumT01-sCoumT13) or the Phe-Tyr route (pCoumPT, sCoumPT01-sCoumPT13) and evaluated in batch and fed-batch culture conditions (Figure 6.3).

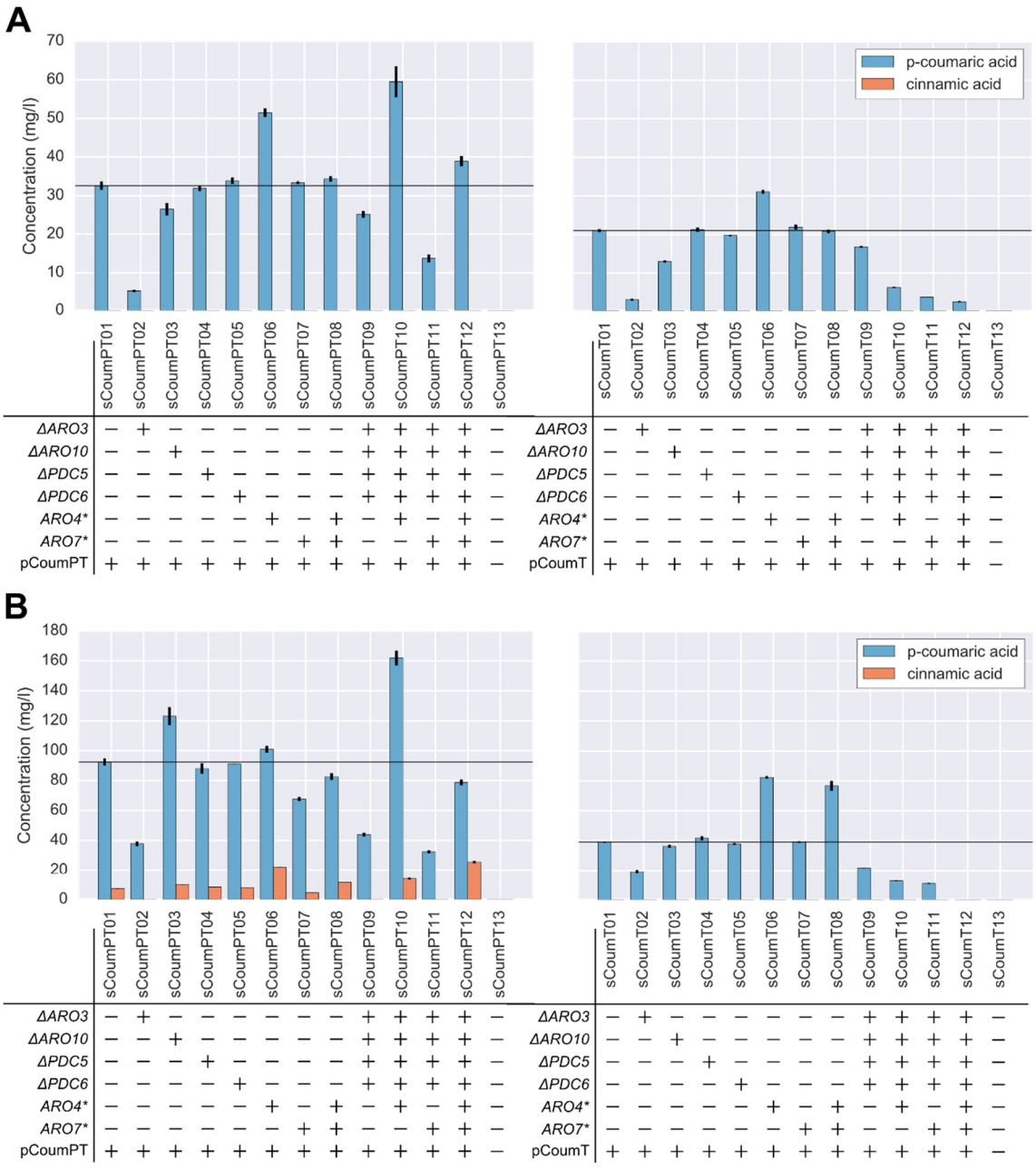

Figure 6.3: Effect of alleviating the feedback inhibition on DAHP synthases or chorismate mutase (*ΔARO3, ARO4** and *ARO7**) and/or deleting (phenyl)-pyruvate decarboxylases (*ΔARO10,*

ΔPDC5 and *ΔPDC6*) on the production of p-coumaric acid. Production titers of strains carrying either pCoumPT (Phe-Tyr route) or pCoumT (Tyr route) when grown in synthetic defined (SD) medium (A, batch) or in Feed-In-Time (FIT) medium (B, fed-batch). Error bars represent the standard error of the mean (n = 3, biological repeats). Strains sCoumPT01 and sCoumT01 were used as reference strains (horizontal black line), sCoumPT13 and sCoumT13 carrying an empty plasmid backbone were the negative control strains. *ARO4** and *ARO7** correspond with *ARO4*^{G226S} and *ARO7*^{G141S}, respectively.

6.4.2.1 Batch versus fed-batch conditions for p-coumaric acid production

The importance of culture conditions for the production of p-coumaric acid was demonstrated by culturing the p-coumaric acid production strains in SD medium, a batch medium, and in Feed-In-Time (FIT) medium, which mimics a fed-batch medium by slowly releasing glucose, causing a linear growth profile. With the exception of sCoumT03 and sCoumT08, and sCoumPT03 and sCoumPT06, the type of medium did not drastically change the p-coumaric acid production landscape of the different strains (Figure 6.3, A versus B). Higher p-coumaric acid titers were obtained for all strains in FIT medium (Figure 6.3 and Supplementary Figure S.4.1). For example, when considering the production strains with the wild-type genetic background, the p-coumaric acid titer increased two – to threefold by just growing the strains under another condition (*e.g.* 32.57 ± 1.09 mg/l for sCoumPT01 in SD medium vs. 92.36 ± 2.37 mg/l in FIT medium). This might suggest that p-coumaric acid is mainly formed during the glucose consumption phase, which is generally the case in FIT medium since glucose is slowly released. These glucose-limited conditions extend the exponential growth phase in which amino acids are essential, subsequently leading to higher p-coumaric acid titers in FIT medium. Once again, this demonstrates that choice and optimization of the microbial cultivation conditions is almost as important as altering the strain metabolic background for establishing an economic feasible microbial production platform [325]. Accumulation of cinnamic acid was observed for the sCoumPT strains in FIT medium but not for the strains in SD medium. This indicates that the balance between AtPAL1p and AtC4Hp/AtATR1p needs further tuning, as not all cinnamic acid is converted to p-coumaric acid. Especially the cytochrome P450 monooxygenase AtC4Hp needing an electron carrier to be functional could be a bottleneck here. Cinnamic acid hydroxylase was indeed determined as a rate-limiting step in the production of p-coumaric acid derived compounds via phenylalanine [5,326]. The expression of an extra P450 reductase besides AtATR1p [267], the creation of a fusion protein of this P450 with its oxidoreductase [327,328], or

the addition of multiple copies of the coding genes could contribute to a reduction of cinnamic acid by-product formation.

6.4.2.2 Tyr route versus Phe-Tyr route for p-coumaric acid production

With the exception of Koopman *et al.* [48] who also investigated the influence of using both tyrosine and phenylalanine as flavonoid precursors in *S. cerevisiae*, the majority of earlier studies used either the phenylalanine [49] or tyrosine precursor pool [5,65,212,297]. In this study, a comparison was made between the commonly used tyrosine (Tyr) route and the phenylalanine-tyrosine (Phe-Tyr) route for the production of p-coumaric acid (Figure 6.3, right versus left).

When looking at the production strains sCoumPT01 and sCoumT01, purely illustrating the effect of both pathways on the usage of the native p-coumaric acid precursor pools, a significantly higher titer was obtained when using the Phe-Tyr route (p-value of 5.58E-4 and 2.45E-5 for SD and FIT medium respectively, Figure 6.3). In addition, considering all production strains, with the exception of strain background 08 in FIT medium, the p-coumaric acid titer increased when using both the phenylalanine and tyrosine pool instead of only tyrosine (Figure 6.3 and Supplementary Figure S.4.1), which is expected since both precursors are pulled away for p-coumaric acid production. Comparable results were obtained by Koopman and coworkers, who examined naringenin production with the use of either phenylalanine or both aromatic amino acids [48].

6.4.2.3 Effect of an enhanced flow to aromatic amino acids on p-coumaric acid production

As reported earlier, the unwanted production of aromatic alcohols, depleting the flux to aromatic amino acids, can be avoided by deleting (phenyl)-pyruvate decarboxylases Aro10p, Pdc5p and Pdc6p [48,65]. For the single knock-outs of *PDC5* and *PDC6* no significant differences with the reference strains occurred (p-values > 0.05), which is consistent with earlier reports [48,65]. For the single knock-out of *ARO10* significantly higher titers were obtained in FIT medium when both the phenylalanine and the tyrosine pool is used for production (sCoumPT03, p = 0.009). This effect is supported by the results of Koopman *et al.* [48] as an extra *ARO10* knock-out in their strains also led to higher titers of p-coumaric acid beside naringenin. On the other hand, no effect is observed when the Tyr route is followed, similar to the *PDC5* and *PDC6* knock-outs, indicating that the tyrosine pool, in contrast to

the phenylalanine pool, is not altered by eliminating (phenyl)-pyruvate decarboxylases. The latter is in contrast to literature where these knock-outs improved p-coumaric acid production from tyrosine [65]. In SD medium the effect of this knock-out is negative (p-values at least < 0.036), and more pronounced when only the Tyr route is used, indicating a different balance of and between the two pools, *i.e.* tyrosine and phenylalanine, under the different culture conditions.

The effect of alleviating feedback inhibition by phenylalanine and tyrosine was investigated by deleting the native *ARO3* [48], or by replacing the native *ARO4* (ChrII 716882-717994) and/or *ARO7* (ChrXVI 674861-675631) in the genome of *S. cerevisiae* by genes coding for a tyrosine-resistant DAHP synthase Aro4p^{G226S} and chorismate mutase Aro7p^{G141S}, respectively [48,64,65,224]. The single knock-out of *ARO3* had a negative effect on the production of p-coumaric acid. For the strains sCoumPT02 and sCoumT02 in both media the final titer significantly dropped more than twice compared to the reference strain (p-values smaller than 10E-4). Also, in the quadruple knock-out strain (background 09), where the relieve of phenylalanine feedback is combined with the elimination of by-product formation, production did not improve compared to the reference strain. As such, the deletion of *ARO3*, if not combined with a modified *ARO4*, has a very adverse effect. This is conceivable as the amino acid biosynthesis is tightly regulated by Gcn4p and DAHP synthase is the gateway to the shikimate pathway [329,330]; an *ARO3* knock-out presumably causes deregulation of fluxes in the shikimate pathway. Strains having only *ARO4^{G226S}* alteration (*i.e.* sCoumPT06 and sCoumT06) produce more p-coumaric acid. However, for the strain using both amino acid pools (sCoumPT06), the positive effect is only significant in batch conditions (p = 2.6E-4). On the other hand, for sCoumT06, expressing only *RcTal1*, the tyrosine feedback-resistant DAHP synthase had a positive impact on production, both in batch and in fed-batch conditions, which is consistent with previous observations [65,66] (Figure 6.3). Due to the G226S amino acid replacement in Aro4p, tyrosine can accumulate without any negative effect on the DAHP synthases, further leading to enhanced p-coumaric acid titers. Relieving the feedback can also lead to higher levels of phenylalanine, which can block DAHP synthesis via Aro3p, but in this case, excess of phenylalanine can be pulled away via the production of phenylacetaldehyde. Just as earlier described [64], the engineered chorismate mutase (Aro7p^{G141S}) did not affect p-coumaric acid production compared to the reference strains (p-values > 0.30), with the exception of sCoumPT07 in FIT medium, which resulted in a lower titer. The combination of Aro7p^{G141S} and Aro4p^{G226S} (sCoumPT08 and sCoumT08)

in fed-batch conditions confirmed the results obtained with Aro4p^{G226S}: a similar positive effect on production is seen only when the Tyr route is used. Still, the double mutants do not perform better than the single Aro4p mutant. In batch conditions, combining Aro7p^{G141S} with Aro4p^{G226S} even eliminated the positive effect of Aro4p^{G226S}. In earlier observations, performed by Gold *et al.* [66], combining both tyrosine feedback negative enzymes improved the production of p-coumaric acid compared to single Aro4p feedback negative strains, which is not observed here. In their study however, a K229L amino acid replacement in Aro4p was performed instead of a G226S substitution (and *ARO10* was additionally deleted). It has been demonstrated that both alterations lead to a tyrosine insensitive DAHP synthase, but the G226S mutation makes Aro4p phenylalanine regulated while Aro4p^{K229L} is unresponsive for both amino acids [224]. In our strains, the nonappearance of the extra beneficial effect which could have been obtained in the double mutants could be explained by the fact that too high levels of phenylalanine can block further increased synthesis of DAHP as both DAHP synthases Aro3p [224] and Aro4p^{G226S} [224] are feedback inhibited by phenylalanine.

When looking at combinations of the aforementioned engineering strategies (backgrounds 09 to 12), a remarkable result is the huge difference between strains sCoumPT10 and sCoumT10, both having an identical genomic background (*ΔARO3 ΔARO10 ΔPDC5 ΔPDC6 ARO4^{G226S}*). While in sCoumPT10 eliminating by-product formation and deleting *ARO3* (sCoumPT06 versus sCoumPT10) clearly has an added value on top of the effect of Aro4p^{G226S} (up to 1.6-fold improvement), in sCoumT10 it has quite the opposite effect (up to 6.5-fold decrease). Strain sCoumPT10 produced in both media at least a 10-fold more p-coumaric acid than sCoumT10 (Figure 6.3, Supplementary Table S.4.4 and Supplementary Table S.4.5). In both strains, by-product formation is strongly reduced due to the deletion of (phenyl)-pyruvate decarboxylases Aro10p, Pdc5p and Pdc6p, leading to enhanced phenylalanine pools. Yet, the remaining Aro4p^{G226S} is allosterically inhibited by phenylalanine [224]. In strain sCoumPT10, carrying the Phe-Tyr route to p-coumaric acid, the excess of phenylalanine can be pulled away toward p-coumaric acid while for sCoumT10 phenylalanine is piling up, which consequently blocks the activity of Aro4p^{G226S} and further reduces the biosynthesis of p-coumaric acid as both production of phenylalanine and tyrosine are halted. Comparing the results to those of the strains not engineered toward phenylacetaldehyde loss (*i.e.* sCoumPT01 versus sCoumPT06, 07 and 08), the results are confirmed in that way that Aro4p^{G226S} has a positive effect (sCoumPT10 versus

sCoumPT09), Aro7p^{G141S} has no or a negative effect (sCoumPT11 versus sCoumPT09), and combining Aro7p^{G141S} with Aro4p^{G226S} largely eliminates the positive effect of Aro4p^{G226S} (sCoumPT12 versus sCoumPT09).

The negative results obtained with sCoumT11 and 12 can be explained in a similar way as for sCoumT10. In these strains with combined engineering strategies using only the Tyr route (strains sCoumT09 versus sCoumT10 to sCoumT12), the results suggest that Aro4p^{G226S} and Aro7p^{G141S} have a significant negative effect on p-coumaric acid production (p-values for DAHP synthase and chorismate mutase, after performing a two-way ANOVA analysis, were respectively 8.61E-10 and 3.96E-11 in SD medium and 2.26E-10 and 6.10E-11 in FIT medium (Figure 6.3, Supplementary Table S.4.6 and Supplementary Table S.4.7)). In this context, always one of the DAHP synthases is feedback-inhibited through either tyrosine or phenylalanine causing presumably these unfavorable production amounts. Again, introducing Aro4p^{K229L}, feedback resistant for both phenylalanine and tyrosine, could be a solution to further enhance production [65,297].

Comparing our highest production titers in fed-batch conditions to those reported earlier [65], the titers in this study are still more than a 12-fold lower (161.91 mg/l vs. 1.93 mg/l). This can be explained by some fundamental differences between our best producer (sCoumPT10) and the strain of Rodriguez *et al.* [65] regarding the strain background, the engineered Aro4p and the used *Tal* gene. Several studies revealed the importance of the *S. cerevisiae* background (*e.g.* S288c vs. CEN.PK) for the construction of cell factories [331–334]. Both the S288c and CEN.PK strains differ in more than 22.000 single nucleotide polymorphisms (SNPs) of which 13.000 are present in 1843 ORFs, possibly leading to changes in protein activity [331]. Indeed, specifically for the biosynthesis of p-coumaric acid, it was shown that in the CEN.PK background 20 to 50% higher p-coumaric acid titers were obtained compared to S288c (*i.e.* the background strain used in this study) [334]. Additionally, their p-coumaric acid production strain also had an extra *E. coli* isoenzyme of the shikimate kinase (AroLp), which led to an extra increase in p-coumaric acid titer from 1.0 mg/l to 1.93 mg/l. In addition, even in similar genetic backgrounds and culture conditions, final p-coumaric acid concentrations can strongly differ. For instance, with the *ΔARO10 ΔPDC5 ARO4^{K229L} ARO7^{G141S} Tal* background grown in FIT medium for 72h, either titers of around 1.0 g/l [65] or 3.28 mg/l.OD [297] were reached. Even with an OD of 100 for the latter, the produced p-coumaric acid amount is far lower than 1.0 g/l. Furthermore, p-coumaric acid

production with the Phe-Tyr route could possibly be further improved by using one of the highly active tyrosine ammonia lyases. It was demonstrated that the Talp of *Flavobacterium johnsoniae* (*FjTal*) or *Herpetosiphon aurantiacus* (*HaTal1*) led to around a 3-fold higher p-coumaric acid titer than our RcTal1p [212]. However, the fluxes in this heterologous production pathway should than also be carefully balanced, as high p-coumaric acid amounts are no guarantee for high naringenin production [49]. In this study, we also worked with a feedback resistant Aro7p^{G141S} and a phenylalanine regulated Aro4p^{G226S} whose native genes were modified in the genome without any overexpression, which is in contrast with most studies overexpressing the DAHP synthase and chorismate mutase at other loci or expression vectors and using the total feedback resistant Aro4p^{K229L}.

Nevertheless, sCoumPT10 (*ΔARO3 ΔARO10 ΔPDC5 ΔPDC6 ARO4^{G226S}* pCoumPT) led to the highest production titer of p-coumaric acid in both SD and FIT medium after 72h of growth, which was 59.50 ± 4.02 mg/l and 161.91 ± 4.90 mg/l, respectively. In both cases this was ca. a 2.0-fold improvement compared to the reference strain with the wild-type genetic background (sCoumPT01). The best p-coumaric acid producer in this study had the same genetic alterations as for the optimized naringenin production strain of Koopman *et al.* [48] and led there to a 3.0-fold improvement of naringenin production. This confirms the importance of removing (phenyl)-pyruvate decarboxylases and alleviating the negative feedback mechanisms of the DAHP synthases to enhance flavonoid production in *S. cerevisiae*. This strain background (*ΔARO3 ΔARO10 ΔPDC5 ΔPDC6 ARO4^{G226S}*), together with the *ARO4^{G226S}* background also leading to high p-coumaric acid titers in batch conditions, were selected for assessing naringenin production.

6.4.3 Effect of an enhanced malonyl-CoA pool on naringenin production

To examine if an enhanced cytosolic malonyl-CoA pool improves the production of naringenin, strains were constructed only carrying the downstream part of the naringenin pathway (pNar) whether or not complemented with the *ACC1^{S659A,S1157A}* overexpression vector (pOEACC1^{S659A,S1157A}). With view on *de novo* naringenin production, this was performed in the wild-type strain SY992, serving as a reference, and the strain backgrounds leading to the highest *de novo* p-coumaric acid titers (Table 6.2, leading to strains sNarC01-sNarC04 and sNarAC01-sNarAC04). Using this strategy, by feeding with p-coumaric acid, solely the influence of malonyl-CoA on naringenin production could be evaluated. For all

these strains, growth was characterized by measuring the endpoint OD600 after 72h which indicated no specific growth deficiencies of the flavonoid production strains.

As expected, significantly higher naringenin production titers were obtained in all strains overexpressing the enhanced acetyl-CoA carboxylase ACC1p^{S659A,S1157A} (p-values by pairwise comparison of sNarC and sNarAC strains were smaller than 4.92E-4). A maximum improvement of up to 2.2-fold was obtained for strain sNarAC03 compared to sNarC03, leading to a final naringenin concentration of 12.96 ± 0.62 mg/l (Figure 6.4, Supplementary Table S.4.8). Probably, native malonyl-CoA concentrations in the yeast metabolism are lower than the K_m of GmCHS5p which is 4.01 µM [335]. As such, these results suggest that extra malonyl-CoA is able to improve the conversion to chalcones via the chalcone synthase GmCHS5p, an enzyme type known to have a rather low catalytic efficiency for malonyl-CoA (k_{cat}/K_m of 15080 $s^{-1}M^{-1}$) [335] compared to its preceding enzyme in the pathway At4CL3p (k_{cat}/K_m of 227900 $s^{-1}M^{-1}$) [336]. Interestingly, the introduction of genetic alterations to increase the common tyrosine and phenylalanine precursor prephenate, *i.e.* like *ARO4^{G226}* and *ΔARO3*, whether or not combined with the deletion of *ARO10, PDC5* and *PDC6* to avoid degradation of phenylpyruvate, immediate precursor of phenylalanine, led to increased naringenin concentrations. Especially for the *ΔARO3 ΔARO10 ΔPDC5 ΔPDC6 ARO4^{G226S}* background strains, a 3.0 and 4.0-fold improvement was observed for sNarC03 (without pOEACC1^{S659A,S1157A}) and sNarAC03 (with pOEACC1^{S659A,S1157A}) compared to sNarC01 and sNarAC01, respectively. It is suggested that a lower flux through the shikimate pathway caused by the negative feedback of accumulating phenylalanine could in these strains improve the flow to pyruvate [66]. Since no better growth was observed for these strains, it is plausible that this promoted the pyruvate to malonyl-CoA conversion. The larger effect in the *ΔARO3 ΔARO10 ΔPDC5 ΔPDC6 ARO4^{G226S}* (sNar(A)C03) compared to the *ARO4^{G226S}* (sNar(A)C02) strains could be attributed to the fact that in the former strains pyruvate cannot be channeled away to acetaldehyde and ethanol due to the *PDC5* and *PDC6* knock-outs. The pyruvate decarboxylase deletions also attenuates the supply of malonyl-CoA, however this supply is further secured in yeast by ScPdb1p and ScPda1p. Remarkably, also some p-coumaric acid consumption was measured in the negative control strains sNarC04 and sNarAC04, without detection of the target metabolites.

With a final titer of 12.96 ± 0.62 mg/l naringenin, the results were in line with earlier studies producing naringenin (12.5 mg/l to 15.6 mg/l) from extracellularly fed p-coumaric acid

[337,338]. Only one study that used the strong inducible *GAL1* promoter in front of every gene reported titers twice as high as ours. Indeed, a lot of potential for further strain improvement is possible since the best production strain sNarAC03 had a naringenin yield of 0.197 ± 0.012 mol mol^{-1} p-coumaric acid which is still a 5.0-fold lower than the theoretical yield of 1.0 mol mol^{-1} p-coumaric acid (Supplementary Table S.4.9). To this end, analysis of C-balances revealed a loss of p-coumaric acid toward unwanted by-product formation, which was indicated as phloretic acid by UPLC-UV analysis (Supplementary Figure S.4.2).

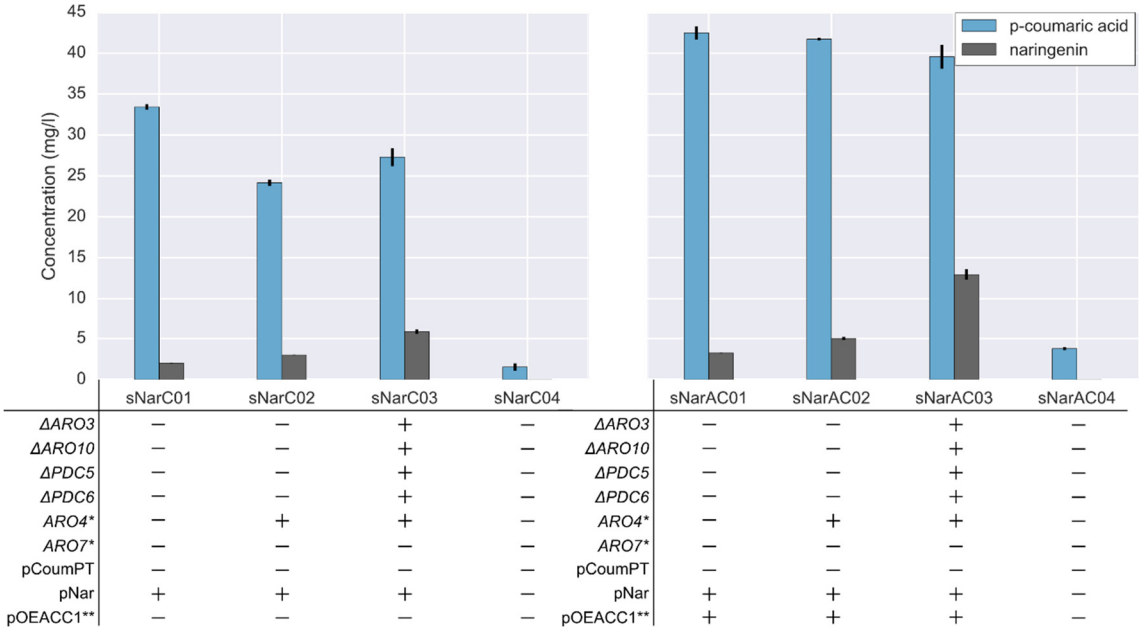

Figure 6.4: Effect of the overexpression of an improved acetyl-CoA carboxylase (ACC1p^{S659A,S1157A}) on the production titers of naringenin in Feed-In-Time (FIT) medium after 72h. A final concentration of 164.05 mg/l (1mM) p-coumaric acid was fed to the production strains. The p-coumaric acid concentrations represent the amount that is metabolized, the naringenin titers represent the amount that is produced. Error bars represent the standard error of the mean (n = 3, biological repeats). Strains sNarC01 and sNarAC01 were used as reference strains, sNarC04 and sNarAC04 were the negative control strains, carrying respectively an empty LEU2, and an empty LEU2 and URA3 plasmid backbone. *ARO4*, ARO7** and pOEACC1** correspond with *ARO4*^{G226S}, *ARO7*^{G141S} and pOEACC1^{S659A,S1157A}, respectively.

6.4.4 De novo production of naringenin: combining enhanced flows toward aromatic amino acids and malonyl-CoA

For naringenin production from glucose, the *At4CL3*, *GmCHS5* and *GmCHI1A* genes expressed on the low-copy vector pNar were introduced in strains sCoumPT01, sCoumPT06 and sCoumPT10 leading to strains sNar01 to 03 and sNarA01 to 03 (Table 6.2). In the sNarA

strains, pOEACC1^{S659A,S1157A} was introduced to obtain an enhanced malonyl-CoA pool. The *de novo* naringenin production capacity of the obtained strains was evaluated in batch and fed-batch medium after 72h. Again, endpoint OD600 was determined as a measure for growth and revealed no strong growth deficiencies.

Modifying Aro4p to a tyrosine resistant DAHP synthase led to an improvement in naringenin titers in all strains and conditions independent of the additional overexpression of ACC1p^{S659A,S1157A}. Moreover, increasing the flow to prephenate by deleting the phenylalanine regulated DAHP synthase Aro3p and increasing the availability of phenylalanine by deleting (phenyl)-pyruvate decarboxylases further improved naringenin titers in the sNarA03 strains by 2.1-fold and 1.6-fold in batch and fed-batch cultivations respectively. This is in line with a similar study where it was shown that the extra deletion of *ARO10* beside the modified Aro4p and ACC1p^{S659A,S1157A} overexpression improved resveratrol biosynthesis with 30% [267]. Yet this improvement is not seen for sNar03 compared to sNar02 which is surprising regarding the results of sNarC03 compared to sNarC02 (section 6.4.3). As such, this indicates an imbalance between the upstream module delivering p-coumaric acid (pCoumPT) and the malonyl-CoA supply in these strains.

Comparable to the results obtained with the p-coumaric acid producing strains, fed-batch conditions led to higher titers of p-coumaric acid and to the additional production of cinnamic acid (Figure 6.5, A versus B and Supplementary Table S.4.10). The naringenin production was equal or lower compared to batch medium. The highest concentrations of naringenin were obtained with strain sNar02 and sNarA03 in batch conditions (4.07 ± 0.24 mg/l and 3.83 ± 0.17 mg/l, respectively). Since glucose is continuously fed at a slow rate in fed-batch conditions, these results again indicate that p-coumaric acid is mainly formed during glucose consumption. Moreover, it could be suggested that the production of naringenin really starts to increase when glucose is depleted. This was also observed in earlier studies for naringenin production in batch fermentations [48] and for resveratrol biosynthesis using a similar pathway [5,267]. Measuring glucose levels in future experiments could give a decisive answer here.

With the introduction of the downstream part of the naringenin pathway (pNar), the biosynthesis of p-coumaric acid was accompanied by phloretic acid production (Figure 6.5). Phloretic acid biosynthesis was also found in other *S. cerevisiae* strains expressing the phenylpropanoid pathway [48,339]. Recently, the endogenous enoyl reductase (ScTsc13p) was

identified as the responsible enzyme for phloretic acid production via reduction of p-coumaroyl-CoA [340]. Since chalcone synthase has a very low catalytic efficiency for p-coumaroyl-CoA (k_{cat}/K_m of 8190 $s^{-1}M^{-1}$) [335], this implies a very efficient conversion of p-coumaroyl-CoA toward phloretic acid via ScTsc13p, which thus strongly competes with GmCHS5p for naringenin production in yeast. Our results show that increasing the malonyl-CoA pool could (partially) solve the loss of p-coumaroyl-CoA toward phloretic acid (Figure 6.4 and Supplementary Figure S.4.2). Indeed, especially when considering sNarA03, higher naringenin amounts and lower phloretic acid titers were observed in both cultivation conditions compared to its native expressing *ACC1* strain sNar03 (*e.g.* naringenin titers: 3.83 ± 0.17 mg/l vs. 1.33 ± 0.91 mg/l in SD medium and 2.92 ± 0.10 mg/l vs. 1.92 ± 0.73 mg/l in FIT medium). Yet, optimizing the malonyl-CoA pool for strain sNar02, resulting in strain sNarA02, indeed lowered phloretic acid concentrations, but in contrast to the positive results obtained for strain sNarA03, this did not result in better naringenin production as only half of the amount or an equal amount was detected in batch and fed-batch cultivations respectively (Figure 6.5). This is remarkable as it was shown in this study (Figure 6.4) and in literature [5,267] that $ACC1p^{S659A,S1157A}$ overexpression combined with a modified Aro4p improved biosynthesis of phenylpropanoid compounds. However, the main difference with our study was the fact that cytochrome P450 reductase (CPR) activity was enhanced with an extra Cyb5p CPR and copy numbers of resveratrol pathway genes were increased. As such, further improvement of naringenin production could be achieved through the introduction of multiple copies of the *GmCHS5* gene, which is known to be a rate-limiting enzyme in flavonoid biosynthesis, and eliminating phloretic acid production since a lot of p-coumaroyl-CoA is pulled away via this route. As a knock-out of the *TSC13* gene is lethal in yeast due to its essential role in the elongation of very long chain fatty acids needed for membrane formation [341], it was recently demonstrated that phloretic acid by-product formation could be completely eliminated by replacement of *TSC13* with plant homologues like *Arabidopsis thaliana* (AtECR), *Gossypium hirsutum* (GhECR2) or *Malus domestica* (MdECR) enoyl-CoA reductases which do not show any activity on p-coumaroyl-CoA [340].

In conclusion, with the most optimized background sNarA03, a 1.7-fold and 7.0 fold improvement of naringenin titers compared to the non-optimized flavonoid precursor strain sNar01 in respectively batch and fed-batch conditions was obtained, reaching 3.83 ± 0.17 mg/l (5.24 ± 0.23 mg/g CDW) and 2.92 ± 0.10 mg/l (4.17 ± 0.14 mg/g CDW) of naringenin after 72h of cultivation at deepwell MTP scale. This is still far away from the

shake flask (54 mg/l) and reactor titers (113 mg/l) obtained by Koopman *et al.* [48], however in fed-batch conditions on MTP, this is ca. a 2.0-fold improvement in naringenin amount compared to a similar study using a *ARO7*^{G141S} *ARO4*^{K229L} *aroL ΔARO10 ΔPDC5* strain optimized in its p-coumaric acid pool via the *Tal* gene of *Flavobacterium johnsoniae* (*FjTal*). More specifically in that study, a high p-coumaric acid producer (1.93 ± 0.26 mg/l [65]) led, after introduction of the pathway genes, to a naringenin titer of only 1.55 ± 0.13 mg/l [49]. Nevertheless, final naringenin concentrations were still lower as for the experiment when p-coumaric acid was fed (Figure 6.4) which indicates unbalanced supply of p-coumaric acid in the upstream part of the pathway. Since we only evaluated in this study genetic alterations to enhance all three flavonoid precursor pools (*i.e.* push strategy), possibly leading to imbalances of metabolite intermediates in the cell, it will be useful in the future to further balance the activity of the enzymes in the naringenin pathway (*i.e.* pull strategy). This could imply the introduction of extra gene copies, preferably by integration in the genome and the complete elimination of phloretic acid by-product formation. Also organizing enzymes in synthetic protein scaffolds or cell organelles, enzyme engineering and multivariate modular metabolic engineering between the up – and downstream module (*e.g.* pCoumPT and pNar), by for example using different (synthetic) promoter – and 5'UTR libraries, are worthwhile strategies to improve *de novo* naringenin biosynthesis in yeast [55]. The latter is especially interesting in this study as transcriptional – and translational control elements can be easily switched in the up – and downstream module by using the VEGAS assembly technique. Finally, as higher naringenin titers were observed in batch conditions, suggesting that naringenin is mainly formed when glucose starts to deplete, decoupling growth and production via dynamic pathway control [211,342] by using for example glucose repressed yeast promoters that are only activated in the production phase [343], could probably also contribute to higher production titers.

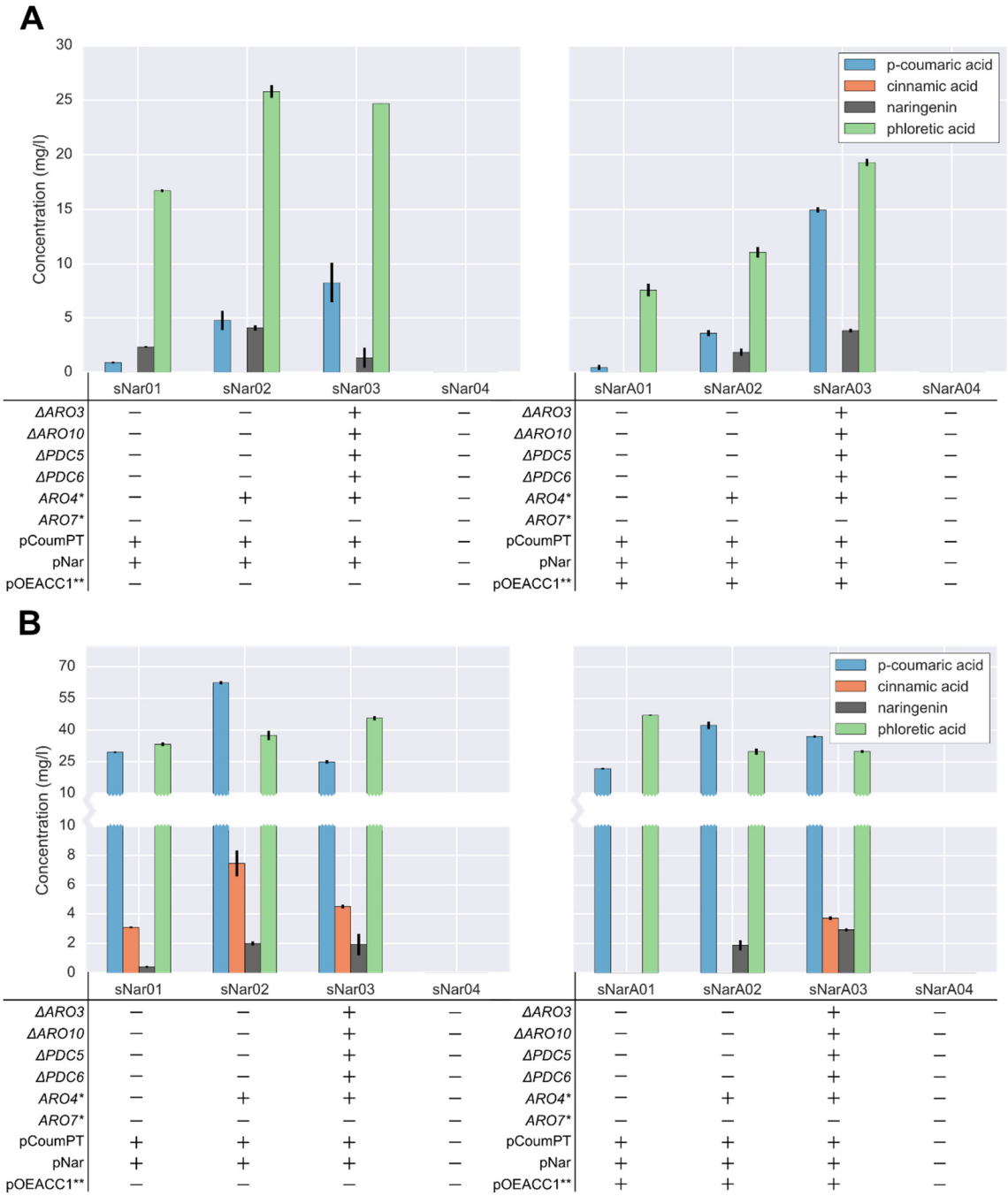

Figure 6.5: *De novo* production of naringenin in yeast strains with an improved pool of p-coumaric acid whether or not completed with an increased malonyl-CoA pool. Strains were grown for 72h in either Synthetic Defined (SD) (A) or Feed-In-Time (FIT) medium (B). Error bars represent the standard error of the mean (n = 3, biological repeats). Strains sNar01 and sNarA01 were used as reference strains, sNar04 and sNarA04 were the negative control strains, carrying respectively an empty LEU2 and empty HIS5 plasmid backbone, and an empty LEU2, empty HIS5 and empty URA3 plasmid backbone. *ARO4*, ARO7** and pOEACC1** correspond with $ARO4^{G226S}$, $ARO7^{G141S}$ and pOEACC1^{S659A,S1157A}, respectively.

6.5 CONCLUSION

The many interesting biological properties of phenylpropanoids make these molecules very interesting targets for the pharmaceutical and agricultural industry. As plant extraction of these compounds leads to low yields and the use of hazardous solvents and harsh reaction conditions, microbial production using renewable carbon sources is an attractive alternative. In this study we simultaneously optimized for the first time all three precursors for the biosynthesis of phenylpropanoid compounds (*i.e.* phenylalanine, tyrosine and malonyl-CoA) in *Saccharomyces cerevisiae*, with the production of the flavonoid naringenin as proof of concept.

First the phenylalanine and tyrosine precursor pools were optimized and evaluated by measuring p-coumaric acid. Using both amino acid precursors, *i.e.* Phe-Tyr route, clearly improved p-coumaric acid production compared to utilizing only tyrosine, *i.e.* Tyr route. This finally led to an optimized p-coumaric acid pool of around 161.91 ± 4.90 mg/l in fed-batch conditions. Second, improving cytosolic malonyl-CoA supply positively affected naringenin production reaching a final titer of 12.96 ± 0.62 mg/l. Combining both approaches finally led to *de novo* naringenin production titers of around 4.0 mg/l in deepwell MTP cultivations which was ca. a 2.0-fold improvement compared to a similar study using only the tyrosine precursor. This study also revealed that the cultivation conditions have a large impact on production efficiency. The highest naringenin titers were observed in batch conditions. In contrast, a more efficient production of p-coumaric acid was obtained in fed-batch conditions.

Nevertheless, a loss of carbon was detected toward the production of phloretic acid indicating the ability here for strain improvement. Additionally, since we only enhanced the flavonoid precursor pools, further optimization of the naringenin pathway itself, by *e.g.* multivariate modular metabolic engineering and protein scaffolds, will be needed in the future to increase production amounts. Especially for the rate-limiting enzymes like C4Hp and CHS5p. Therefore, this naringenin production strain with already optimized precursor pools forms a valuable starting point for further development of a naringenin yeast cell factory.

CHAPTER 7 GENERAL DISCUSSION AND OUTLOOK

In recent years, the yeast synthetic biology toolbox rapidly expanded, enabling yeast cell factory development for the efficient production of fine and bulk chemicals, and even leading to the construction of the world's first fully synthetic yeast genome (*i.e.* Sc2.0 Consortium). This toolbox typically exists of modular regulatory parts (*e.g.* promoters, terminators, transcription-based sensors, *etc.*), spatial engineering tools like protein scaffolds and tags for localization of pathways in yeast organelles, and genome editing tools such as Yeast Oligo-mediated Genome Engineering (YOGE), Transcription activator-like effector nucleases (TALENs), zinc-finger nucleases (ZFNs) or CRISPR/Cas9 [28]. Despite the great effectiveness of these yeast synthetic biology tools, the field is still hampered by the lack of well-characterized (synthetic) modular parts and predictive methods for reliable strain development. Enhancing the predictability of the design-build-test cycle is indispensable to speed up the construction process of economically feasible yeast production hosts. To this end, the main objective of this doctoral research was to broaden the synthetic biology toolbox to facilitate and fasten the construction of *Saccharomyces cerevisiae* cell factories. As such, different engineering techniques for modulating gene expression at the transcriptional and translational level were developed and evaluated for their ability, reliability and predictability in altering the cell's metabolism (Figure 1.1). Moreover, to assess the potential of *S. cerevisiae* as an industrial host for the production of secondary metabolites, this yeast was metabolically engineered for the production of flavonoids as a proof of concept (Figure 1.1). Therefore, advanced synthetic biology tools like CRISPR/Cas9 and the versatile genetic assembly system (VEGAS) were used and one of our developed methods was for the first time evaluated on a flavonoid pathway gene leading to predictable p-coumaric acid production (**Chapter 4**).

Evaluation of standardization approaches in the synthetic biology field

To really improve the predictability of engineering biological systems and transforming the synthetic biology field to a mature engineering discipline, extensive standardization and sharing of research (meta)-data is needed. In **Chapter 2**, a critical evaluation of the efforts already undertaken, and the lack thereof, regarding standardization in all stages of the strain development cycle was performed. In comparison to well-established engineering disciplines like electronics, using standardized parts for electronic circuit design, setting up such a policy in the field of synthetic biology is still in its infancy. Though part repositories

like the iGEM Registry of Standard Biological Parts and the Joint BioEnergy Institute's Inventory of Composable Elements (JBEI-ICE) have been set up, the main challenge for biological parts remains linking their physical demarcation with a desired predictable functional performance. This is especially difficult considering the influence of other regulatory elements in the immense complex architecture of a cell. In addition, the lack of uniform protocols for part characterization contributes to a huge variety of different measures for a part's performance. Together with the diverse ways of how info of characterized parts is shared (*e.g.* FASTA, SBML, *etc.*), this impedes the reusability of evaluated parts by the community. In view of the growing importance of reliable strain engineering and the associated usage of computer-aided design (CAD) software, sharing of data on generated and characterized parts should be completely standardized. To this end, uniform and standardized exchange formats, metadata and units are essential to unambiguously interpret biological part data. Specifically, such multi-omics data sets of regulatory parts are necessary for data-driven methods, *e.g.* machine learning techniques, enabling forward engineering of biological systems. With an increase of adequate uniform reported data, these models can be further improved and trained, as such enhancing their accuracy and thus reducing the number of design candidates. Altogether, this will lead to faster strain development and an overall reduction in costs of establishing a profitable production process.

Development of tools to expand the yeast synthetic biology toolbox

In order to contribute to the improvement of the design-build-test cycle, especially for eukaryotic hosts, novel tools for the engineering of *S. cerevisiae* were developed. The aim was to focus on i) the creation of short (synthetic) parts for transcriptional regulation, ii) the *de novo* development of 5' untranslated regions (5'UTRs) with a predictable effect on a gene's translation initiation and iii) the evaluation of multicistronic expression enabling the reduction of regulatory parts, as such avoiding the repetitive use of biological parts.

As previously mentioned, characterized biological parts become increasingly important for the further establishment of synthetic biology as a real engineering discipline. Specifically, the yeast synthetic biology field is somewhat lagging behind regarding well-characterized regulatory parts. This is mainly due to the complexity of the eukaryotic transcriptional and translational machinery, respectively lacking for example a fixed consensus transcription start site and Kozak sequence, the eukaryotic counterpart of the Shine-Dalgarno sequence.

As such, it is often very difficult to fully demarcate a eukaryotic biological part with reliable performance. Therefore, in **Chapter 3**, the minimal length of the well-characterized native *TEF1* core promoter was elucidated by cumulative truncation. Next, this *TEF1* core promoter served as template for the construction and characterization of a set of short semi-synthetic eukaryotic promoters having a length of only 69 bp. This set of yeast core promoters displayed a 4.0-fold expression range with a maximal promoter activity twice as strong as some long native yeast promoters, even without the need of upstream activating sequence (UAS) elements. Yet, this expression range could be further extended when the core promoter was preceded by one or multiple native UASs. The latter is also described by Redden *et al.* [111], who combined fully synthetic core promoters and UASs to lower the chance on strain instability through unwanted homologous recombination. The short core promoters designed and characterized during this Ph.D. dissertation can be easily incorporated in synthetic oligo's facilitating the assembly of transcription units. Despite this huge benefit, the usage of minimal promoters is still not a common practice in yeast pathway development. While regular native yeast promoters have proven their usability in establishing whole production pathways, minimal promoters are up-to-date only validated in front of the yECitrine reporter. Our results consolidate the potential of minimal yeast promoters in reliable and easy engineering of yeast cell factories [111]. Yet, an interesting future perspective could be to characterize these parts in front of totally different reporters or pathway genes, as such promoting their broad usefulness for yeast synthetic biology. Also genomic stability studies to determine the chance of recombination would be interesting, especially due to the fact that our semi-synthetic promoters still have a native part of around 50 bp which could be sufficient to trigger homologous recombination.

Keeping in mind the importance of reliable and predictable DNA parts and the need for novel (synthetic) regulators, 5'UTR sequences, which play a decisive role in an mRNA translation, were constructed in **Chapter 4**. The great advantage of expression regulation via RNA is its exceptional programmability, making it an attractive molecule for predictive part development. As such, RNA technology is emerging in the synthetic biology field for modulating gene expression, building genetic circuits, detecting molecules, reporting cellular processes and building nanostructures [25]. In case of regulating gene expression, varying the translation initiation rate by modification of the 5'UTR has been proven to work well. Especially in prokaryotes this led to the predictive design of ribosome binding site (RBS) sequences [30,181]. Regarding eukaryotes, altering a gene's translation is generally

overlooked, as mostly no distinction is made between the promoter as a pure transcriptional regulator and the 5'UTR with the Kozak sequence as a pure translational regulator. Consequently, this technology for altering a gene's translation in yeast is still in its infancy and needs more attention to demonstrate its potential in eukaryotes [23,239]. To this end, the yUTR calculator, an *S. cerevisiae* counterpart of the bacterial RBS calculator, was developed and evaluated. Based on an existing data set linking yeast 5'UTR sequences with protein abundances [23], a partial least square (PLS) regression model was built, linking 13 features of the 5'UTR with the outcome of protein expression. The model showed good predictability on an independent test set and was further used for the *de novo* design of reliable 5'UTRs under diverse contexts of promoters, 5'UTRs and coding sequences. In all cases, adequate coefficients of determination (R^2) were achieved, indicating the general applicability of the developed PLS model for forward engineering purposes in yeast. Furthermore, to evaluate the adaptability of the yUTR calculator beyond the use of fluorescent reporters, *de novo* 5'UTRs were developed for the *Rhodobacter capsulatus Tal1* (*RcTal1*) coding sequence and tested for their predictable effect on p-coumaric acid production (**Chapter 4**). The obtained results were promising in that way that effectively 5'UTRs with weak and 5'UTRs with high predicted protein abundances led to lower and higher production titers of p-coumaric acid, respectively.

To the best of our knowledge, we here describe for the first time in *S. cerevisiae* an approach to design *de novo* 5'UTRs with a predictive outcome and generally applicable in different genetic contexts, even on a pathway gene for secondary metabolite production. Nevertheless, since the used features were only determined by randomizing the 10 nucleotides in front of the start codon and yeast 5'UTRs are in general much longer, with an average of 83 bp, expanding the model with sequence and structural features of an entire 5'UTR deserves further investigation. To this end, novel interesting features were recently determined out of a library of half a million 50-nucleotide-long 5'UTRs [239]. In addition, breakthroughs in the structure-activity relationship of yeast core promoters revealed the possibility for programmable engineering of the transcriptional machinery [196]. For instance, a consensus TATA box and an adenine rich initiation region caused higher promoter activities. Combining both approaches in one general applicable model to engineer transcriptional and translational regulation in one step could lead to highly accurate construction and optimization of (heterologous) pathways. Our study together with the pioneering work of Dvir and coworkers [23] and the recent convolutional neural network

(CNN) model of Cuperus *et al.* [239] are good starting points and show the promising, rather unexplored possibility to adjust the translation initiation rate in yeast as a tool for reliable pathway balancing.

Another interesting technique for eukaryotic pathway balancing at the translational level are 2A peptides, enabling multicistronic expression and yielding equimolar amounts of proteins. Regarding modular metabolic engineering, enzymes with similar activities could as such be combined in one multicistronic pathway module. In addition, with the ingrained usage of native promoters and terminators in yeast, the number of regulatory sequences is strongly reduced, which limits the reuse of these parts for long pathways and as such avoids unwanted homologous recombination. While polycistronic expression in eukaryotes is well-established in plant biotechnology and medicinal research, its use in industrial biotechnology is rather scarce, especially in *S. cerevisiae*. Only two studies describe the usage of bicistronic constructs [35,289] and only one study was able to successfully produce β-carotene in *S. cerevisiae* via tricistronic expression [288], for the rest, the use of multicistronic expression in *S. cerevisiae* for other (secondary) metabolite production is nil. **Chapter 5** aimed to evaluate the effectiveness of 2A peptides for metabolic engineering in *S. cerevisiae*, as this was never described before. Novel 2A peptide sequences of the *Thosea asigna* virus were designed on nucleotide level and combined in transcription units with up to four different coding sequences, which all were integrated in the genome by CRISPR/Cas9. Three important aspects were considered: stable integration in the genome, splicing efficiency of the 2A peptides and gene expression. The developed 2A peptides, with the exception of one T2A with some lower reliability, effectively led to spliced proteins and no homologous recombination between 2A sequences was observed. Especially the latter is a risk in polycistronic expression when using multiple 2A peptides in one construct given the huge similarities in nucleotide sequence. Yet, the expression of proteins did not entirely meet the desired efficiencies. While bicistronic transcription units still gave acceptable protein levels, these levels significantly dropped in tri – or quadcistronic expression units. In addition, it remained unclear to which extent the total observed expression was caused by cleaved and/or uncleaved proteins. As such, the unreliable outcome of gene expression, especially in long transcription units, is the main disadvantage of this approach and could explain the current limited use of this synthetic biology tool in *S. cerevisiae* strain engineering. In the future, it could therefore be useful to test the introduction of bidirectional promoters

between short, multicistronic expression units (two to maximum three CDSs) to stimulate the usage of this tool in eukaryotic pathway design.

All three synthetic biology tools studied, created or evaluated during this Ph.D. dissertation, namely the search for a minimal yeast core promoter, the development of the yUTR calculator enabling forward engineering of 5'UTRs and the assessment of the potential or ineptness of 2A peptides for multicistronic expression in *S. cerevisiae* were initially intended to improve standardization and predictability in yeast engineering, and to improve genetic stability of the engineered strains. However, the unreliable gene expression of long constructs with 2A peptides (**Chapter 5**) makes that with this tool mostly a combinatorial trial-and-error approach needs to be followed. As such, with this synthetic biology tool a (high-throughput) evaluation is always needed to tune the ratio between the pathway enzymes. With the ever increasing complexity of programming yeast cell factories, engineering methods based on trial-and-error become too cumbersome. Therefore, tools for reliable forward engineering are indispensable, indicating that our yUTR calculator (**Chapter 4**), which designs *de novo* 5'UTR sequences with a predictable outcome, has a greater potential in the field of yeast synthetic biology. The model developed for this yUTR calculator is based on modifying RNA to predictably alter a gene's translation initiation rate. In contrast to protein parts, where the interaction with other proteins or DNA is difficult to predict, the structure of RNA is strongly related to its function, making it an excellent molecule for the development of *de novo* regulatory parts with user-defined functions. Indeed, the yUTR calculator proved to be useful for the *de novo* design of 5'UTRs with a predictive outcome under diverse contexts of promoters, 5'UTRs and coding sequences, even for a pathway gene applied in secondary metabolite production. Combining this approach for engineering translational regulation with one for transcriptional engineering (*e.g.* Lubliner *et al.* [196]) in one generally applicable model could lead to highly accurate construction and optimization of (heterologous) pathways. However, the complex yeast promoter structure (*e.g.* nucleosomes, upstream regulatory sequences, *etc.*) and the fact that its function is typically based on protein-DNA interactions which are hard to predict, explain why this is today very challenging to aim for. In this respect, during this doctoral research, a range of short yeast promoters was developed, which can be easily incorporated in synthetic oligo's as they are only 69 bps long, facilitating the assembly of transcription units (**Chapter 3**). Our results consolidate the potential of minimal yeast promoters, and given their shortness, these minimal promoters have high potential for the reliable and easy

engineering of yeast cell factories. To promote their broad usefulness for yeast synthetic biology, these parts need to be characterized in front of different reporters and pathway genes. Such characterization will in the end also enable their use in transcription-based forward engineering tools for yeast, and in combined transcription – and translation-based forward engineering tools, as proposed above.

Such computational methods are indispensable tools for the design of biological systems with a predictive outcome. The use of data-driven approaches, like for example machine learning methods, have the big advantage that the search space for appropriate candidates of parts, biological circuits or strains is seriously narrowed, enabling faster construction of cell factories with a desired behavior. However, to develop models with a general applicability, huge amounts of data are needed since the accurate predictive performance of these algorithms strongly depends on the training data sets used for learning the model. The generation of these tremendous amounts of data has been made possible today by the ever increasing innovation in high-throughput facilities such as DNA synthesis, next-generation sequencing, cell picking and pipetting robots, microfluidics and flow-cytometry. As a result, the main challenge recently arising in the field is analyzing and restructuring these data in a well-considered manner usable for computational methods. In this respect, databases with thoroughly standardized info about biological parts, production pathways and even whole strains are still necessary to promote the further development of reliable forward engineering of microbial cell factories (**Chapter 2**), and not the least for *S. cerevisiae*.

Construction of a yeast cell factory for flavonoid production

With the main goal of industrial biotechnology being the development of sustainable processes for the production of economically relevant compounds, the construction of an *S. cerevisiae* cell factory for the biosynthesis of flavonoids, and more specifically naringenin, is an interesting proof of concept. An evaluation of several yeast cell factories was described in **Chapter 6**. The development of production strains needs a combination of metabolic engineering and synthetic biology tools to achieve profitable production titers. In this respect, the metabolism of an S288c *S. cerevisiae* derived reference strain was modified to enhance the supply of precursor molecules for flavonoid production, *i.e.* phenylalanine, tyrosine and malonyl-CoA. Rewiring the metabolism is in recent years facilitated by very efficient genome editing tools such as CRISPR/Cas9, which has proven to be very effective

in eukaryotes and especially in *S. cerevisiae* [19]. One of the great advantages of this technique in yeast metabolic engineering is the possibility of marker-less integration of mutations, knock-outs and even whole pathways at different loci in the genome in one single step. Hence, mutations in the native genes for altering the phenylalanine and tyrosine metabolism were introduced via CRISPR/Cas9. The influence of altering the yeast's metabolism was assessed by the introduction of the naringenin biosynthetic pathway, which exists of seven genes, plant and bacterial derived. Taking into account that for reliable expression in yeast every coding sequence needs a promoter and a terminator, the use of a modular, standardized assembly method was crucial here. Therefore, the versatile genetic assembly system (VEGAS) from the Boeke lab [21] was used. This two-step assembly approach exploits on the one hand the big advantage of Golden Gate, enabling the modular assembly of biological parts into transcription units, and on the other hand the efficient yeast homologous recombination capability, enabling the subsequent connection of the transcription units into whole pathways. Indeed, for the assembly of larger DNA constructs (*i.e.* multiple kilobases to even megabases), and since synthetic gene fragments are still limited to lengths of around 2 – 3 kilobases [344], *in vivo* assembly by using the homologous recombination machinery of *S. cerevisiae* is the better alternative, as the accuracy of efficient *in vitro* techniques for assembly, such as Gibson assembly, Golden Gate and circular polymerase extension cloning (CPEC) to name the most eminent examples, strongly decreases when assemblies longer than 10 – 15 kb are required. The potential of *S. cerevisiae* as a chassis for DNA assembly is clearly demonstrated in the Sc2.0 project where completely new synthetic yeast chromosomes are built [345,346]. Furthermore, in combination with CRISPR/Cas9, this method promotes the quick assembly of large biosynthetic pathways at specific loci on the yeast genome [297]. As such, it hardly needs saying that both VEGAS and CRISPR/Cas9 are cutting-edge tools for metabolic engineering and pathway construction, generally contributing to immensely decreased development times of novel production strains. Their excellent performance was acknowledged in this doctoral research by the efficient construction of a naringenin yeast cell factory.

The profound evaluation of the influence of different metabolic strain backgrounds on metabolite titers finally led to a strain producing up to 13.0 mg/l naringenin from p-coumaric acid and 4.0 mg/l naringenin *de novo* from glucose. With the exception of one study using galactose inducible promoters, comparable titers of naringenin were achieved as previously reported when p-coumaric acid was extracellularly fed (Table 7.1). For *de*

novo naringenin production, there was almost a 3.0-fold improvement compared to a similar study on deepwell MTP scale which only used tyrosine as improved precursor [49]. Yet, two other published studies reported far higher naringenin concentrations in batch conditions (Table 7.1). Overall, production titers of *de novo* flavonoid production from glucose in yeast are barely higher than 100 mg/l [45,46] (Table 1.1). As these titers are still too low to initiate an industrial production process, there is need for further improvements, not in particular in our production strain. To this end, improvements can be applied in two different stages when developing an industrial bioprocess, *i.e.* engineering the strain itself and optimizing the parameters of the fermentation process [347].

Table 7.1: Comparison of published naringenin titers obtained after production in *Saccharomyces cerevisiae*. If the pathway genes were under control of galactose inducible promoters, this is indicated by GAL. Fed-batch experiments were performed in the Feed-In-Time medium of M2P-Biolabs (Baesweiler, Germany).

Precursor	Final titer (mg/l)	Strain	Cultivation type	Reference
p-coumaric acid				
	15.6	S288c	Batch (shake flask)	[337]
	28.3	S288c	Batch (shake flask, GAL)	[326]
	12.5	CEN.PK	Batch (shake flask)	[338]
	13.0	S288c	Fed-batch (MTP)	This study
Glucose				
	54.0	CEN.PK	Batch (shake flask)	[48]
	7.0	S288c	Batch (shake flask)	[47]
	1.6	CEN.PK	Fed-batch (MTP)	[49]
	84.0	S288c	Batch (shake flask, GAL)	[50]
	4.0	S288c	Batch (MTP)	This study

An important factor in the strain engineering process seemed to be the strain type. As in this dissertation an S288c derived yeast strain was used and it was proven that CEN.PK strains are more suitable for p-coumaric acid production [334], and probably thus subsequent phenylpropanoid biosynthesis, it could be interesting to assess flavonoid production in a CEN.PK background. On the other hand, only the supply of the three naringenin precursors was metabolically engineered in this research and no balancing of the pathway itself was executed thus far. In this view, it was indicated in literature that naringenin is mostly formed in the second growth phase, after the consumption of glucose [48]. As such, decoupling growth and production via the use of for instance galactose promoters looks a worthwhile alternative to enhance production [50]. As such, in a first growth phase mainly biomass can be produced while in the second growth phase, when glucose is limited, most cellular energy

and resources can be used for flavonoid production. This approach, together with an improved tyrosine pool, led already to 84 mg/l naringenin on shake flask in an S288c strain (Table 7.1).

Instead of working with inducible promoters, another possibility to optimize the balance between the enzymes in the flavonoid pathway is to specifically regulate and fine-tune every gene's transcription and translation on a constitutive way. In this respect, the developed semi-synthetic promoters and the model-based approach for *de novo* 5'UTR design are useful tools. Especially with the view of how easily *S. cerevisiae* performs homologous recombination, the synthetic promoters (**Chapter 3**) could be the better alternative to avoid strain instability in industrial fermentations. Additionally, in a first strain optimization round, strong and weak synthetic promoters could be placed in front of respectively weakly and strongly active pathway enzymes. Since always remaining pathway intermediates were detected (**Chapter 6**), stronger synthetic promoters could for example be used in front of the downstream genes (*e.g. GmCHS5* and *GmCHI1A*) and weaker synthetic promoters in front of the upstream genes, especially for *AtPAL1*, *RcTal1* and *At4CL3* as not always all cinnamic acid and p-coumaric acid was metabolized. As such, a more continuous flux of intermediates is expected toward the downstream flavonoids. In the same view, the yUTR calculator (**Chapter 4**) for altering translational initiation rates is very useful. Its potential was already shown in a preliminary experiment to predictably vary p-coumaric acid production by modifying the *RcTal1* translation [274]. Compared to the usage of semi-synthetic promoters, this tool can specifically design a 5'UTR with a desired strength for every single flavonoid gene. Since a small set of 5'UTRs with broadly variable strengths is generated, this allows very precise fine-tuning of the flavonoid pathway. Such data-driven tools which enable forward engineering are scarcely used for the moment in *S. cerevisiae* strain engineering and thus could pave the way for more reliable pathway building and optimization, and not only for flavonoids but also for other secondary metabolite production. Furthermore, the yUTR calculator approach could even be expanded, with the generation and evaluation of new data, toward other eukaryotic 5'UTRs, as for example plants. This switch from trial-and-error to more predictable pathway design is currently observed in the field and is driven by the recent developments in CAD software for biological purposes [92,183,184,186]. Such an approach makes it possible to design biosynthetic pathways and genetic circuits with prescribed functions by developing *de novo* regulatory parts (*e.g.* 5'UTRs with predicted behavior) or by using existing biological elements with

characterized properties described in databases (*e.g.* iGEM registry). Ultimately, such CAD software could be used to design the flavonoid pathway by combining our semi-synthetic promoters (**Chapter 3**) with the sets of 5'UTRs developed for every flavonoid gene (**Chapter 4**). From then on, a more semi-combinatorial approach, which preferably requires high-throughput machinery like cloning robots, has to be followed since many combinations of different regulatory parts in the pathway are possible. From a practical point of view, the implementation in the lab of these *in silico* CAD-based biological circuits is facilitated by the modularity of the VEGAS system used in this thesis (**Chapter 6**) which makes it very easy to quickly rearrange (novel) regulatory parts in a pathway.

As every P450 enzyme is accompagnied by a cytochrome P450 reductase (CPR) to ensure a proper electron transfer, it might be worthwhile to express both enzymes in equimolar amounts, which is possible with 2A peptides. Although the 2A peptide approach for tri – and quadcistronic gene expression did not meet our expectations in view of splicing efficiencies and gene expression, bicistronic gene expression works fine (**Chapter 5**) and could therefore be a sensible option to assess the co-expression of *AtC4H* and its CPR *AtATR1*. As such, possibly a better conversion of cinnamic acid to p-coumaric acid, which was seen in **Chapter 6** to be rate limiting, is achievable. Furthermore, carbon is lost via the synthesis of phloretic acid, as such another important step to enhance flavonoid production will be to eliminate this by-product formation through replacement of the native Tsc13p enzyme by a plant homologue [340], which could be achieved by the easy to use and efficient genome editing tool CRISPR/Cas9 [19,298].

The flavonoid pathway genes in this dissertation were also expressed from low-copy vectors. While plasmid-based expression was an easy way here to quickly construct different strains to assess the influence of different metabolic modifications on naringenin production, it is not recommended for the construction of robust industrial strains. Plasmids have some inherent disadvantages like high variation in copy numbers and the need for defined media or antibiotics because of the respectively auxotrophic or antibiotic resistance genes required to maintain the plasmids in the cell. In general and from an industrial point of view, it should be considered to integrate production pathways into the yeast genome which can be very easily implemented with combined CRISPR/Cas9 and VEGAS (see above). In addition, it can be chosen to integrate the pathway on one site or to split up the pathway in different modules and integrate these at different loci. This approach

also allows to incorporate multiple copies of pathway genes or modules, which could be desirable to further enhance production titers. Yet, the chromosomal integration sites have to be carefully chosen as transcription levels can vary at different places in the genome [120]. While 20 well-characterized integration sites are already elucidated in yeast [120], further research in these fundamental properties of the *S. cerevisiae* genome is necessary.

Next to strain engineering efforts, fine-tuning of the bio-process itself should be considered as well to improve production titers. Koopman *et al.* for instance showed that scaling-up from shake flask to batch bioreactors increased the naringenin titer more than twice [48]. In addition, resveratrol titers were doubled by switching from batch to fed-batch bioreactors [5]. To this end, changing from deepwell-MTP (**Chapter 6**) to shake flask and finally (fed)-batch bioreactors should be an effective way to further improve our naringenin titer toward economically feasible quantities. Furthermore, process optimization parameters like aeration, feed rates and medium composition could also contribute to higher product titers. Beside, fermentation in bioreactors enables us to reveal possible strain instabilities which are frequently not observed on deepwell-MTP or shake flask scale [348] and absolutely must be avoided when switching to a real industrial, high-volume bioprocess.

In conclusion, the production strains generated in **Chapter 6** were optimized in their precursor pools for phenylpropanoid biosynthesis and are as such a valuable starting point for further strain and process engineering, which will be needed to fully exploit the potential of *S. cerevisiae* as a robust production host for flavonoid production. In this view, and especially for the further enhancement of the naringenin pathway, the tools developed in this Ph.D. dissertation will be of great aid.

General conclusion

With the ever increasing complexity of production pathways implemented in microbial hosts, the construction of economically feasible *S. cerevisiae* cell factories and microbial cell factories in general still remains a laborious task. To this end, a first scope of this doctoral research was the development of novel synthetic biology tools to facilitate the harmonization of (heterologous) pathways in yeast. Several techniques able to modify gene expression at the transcriptional and translational level were created and evaluated. While all tools showed promising results for altering eukaryotic production pathways, the yUTR calculator (**Chapter 4**) enabling forward engineering of 5'UTRs could be seen as a

breakthrough for the yeast synthetic biology field, especially since predictable design of biological systems is upcoming. A second goal in this Ph.D. dissertation was the application of synthetic biology tools for the development of a *S. cerevisiae* cell factory. Since secondary plant metabolites have lot of interesting biological properties for human health, *de novo* production of naringenin from glucose was chosen as proof of concept. By using enabling genome editing and assembly tools, the *S. cerevisiae* wild-type was successfully transformed into a cell factory for naringenin production. Yet, further optimization of the heterologous pathway will be needed. In this respect, the tools developed in this doctoral research could be of great support for future rational or combinatorial pathway balancing in *Saccharomyces cerevisiae*, generally accelerating the development times of profitable, sustainable bioprocesses.

APPENDICES

S.1 APPENDIX CHAPTER 3 167
S.2 APPENDIX CHAPTER 4 185
S.3 APPENDIX CHAPTER 5 211
S.4 APPENDIX CHAPTER 6 223

S.1 APPENDIX CHAPTER 3

Table S.1.1: Strains used in this study. All strains are obtained from strain SY992. In all plasmid and strain names, the *TEF1* promoter is shortly named as TEF and the core promoter as cpTEF. The genotype of the plasmids is listed in Supplementary Table S.1.2.

Strain	Genotype/Plasmid	Reference
SY992	*Matα, ura3Δ0, his3Δ1, leu2Δ0, trp1-63, ade2Δ0, lys2Δ0, ADE8*	Euroscarf [205]
sRef-pTEF1	pRef-pTEF1	This study
sRef-pADH1	pRef-pADH1	This study
sRef-pCYC1	pRef-pCYC1	This study
sRef-pPGK1	pRef-pPGK1	This study
sRef-pTDH3	pRef-pTDH3	This study
sRef-bl	p2a_empty	This study
s_UAS-cpTEF_1	p_UAS-cpTEF_1	This study
s_UAS-cpTEF_2	p_UAS-cpTEF_2	This study
s_UAS-cpTEF_3	p_UAS-cpTEF_3	This study
s_UAS-cpTEF_4	p_UAS-cpTEF_4	This study
s_UAS-cpTEF_5	p_UAS-cpTEF_5	This study
s_UAS-cpTEF_6	p_UAS-cpTEF_6	This study
s_UAS-cpTEF_7	p_UAS-cpTEF_7	This study
s_UAS-cpTEF_8	p_UAS-cpTEF_8	This study
s_UAS-cpTEF_9	p_UAS-cpTEF_9	This study
s_cpTEF_1	p_cpTEF_1	This study
s_cpTEF_2	p_cpTEF_2	This study
s_cpTEF_3	p_cpTEF_3	This study
s_cpTEF_4	p_cpTEF_4	This study
s_cpTEF_5	p_cpTEF_5	This study
s_cpTEF_6	p_cpTEF_6	This study
s_cpTEF_7	p_cpTEF_7	This study
s_cpTEF_8	p_cpTEF_8	This study
s_cpTEF_9	p_cpTEF_9	This study
s_cpTEF_6-libA	p_cpTEF_6-libA	This study
s_cpTEF_6-libB	p_cpTEF_6-libB	This study
s_cpTEF_6-libC	p_cpTEF_6-libC	This study
s_cpTEF_6-libD	p_cpTEF_6-libD	This study
s_cpTEF_6-A	p_cpTEF_6-A	This study
s_cpTEF_6-B	p_cpTEF_6-B	This study
s_cpTEF_6-C	p_cpTEF_6-C	This study
s_cpTEF_6-D	p_cpTEF_6-D	This study
s_cpTEF_6-E	p_cpTEF_6-E	This study
s_cpTEF_6-F	p_cpTEF_6-F	This study
s_cpTEF_6-G	p_cpTEF_6-G	This study
s_UAS$_{TEF1}$-1x	p_UAS$_{TEF1}$-1x	This study
s_UAS$_{TEF1}$-2x	p_UAS$_{TEF1}$-2x	This study
s_UAS$_{TEF1}$-3x	p_UAS$_{TEF1}$-3x	This study
s_UAS$_{TEF1}$-4x	p_UAS$_{TEF1}$-4x	This study
s_UAS$_{CLB2}$-1x	p_UAS$_{CLB2}$-1x	This study

s_UAS$_{CLB2}$-2x	p_UAS$_{CLB2}$-2x	This study
s_UAS$_{CLB2}$-3x	p_UAS$_{CLB2}$-3x	This study
s_UAS$_{CIT1}$-1x	p_UAS$_{CIT1}$-1x	This study
s_UAS$_{CIT1}$-2x	p_UAS$_{CIT1}$-2x	This study
s_UAS$_{CIT1}$-3x	p_UAS$_{CIT1}$-3x	This study
s_UAS$_{CIT1\text{-}CLB2}$	p_UAS$_{CIT1\text{-}CLB2}$	This study
s_UAS$_{CIT1\text{-}TEF1\text{-}CLB2}$	p_UAS$_{CIT1\text{-}TEF1\text{-}CLB2}$	This study

Table S.1.2: Plasmids used in this study. All 5'UTRs are the native 5'UTR sequences of the preceding promoter. In all plasmid names, the *TEF1* promoter is shortly named as TEF and the core promoter as cpTEF.

Plasmid	Genotype/Description	Reference
pKT140	yECitrine-tADH1, *KAN*, AmpR, CEN/ARS	Euroscarf[206]
pRef-pTEF1	pTEF1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
pRef-pADH1	pADH1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
pRef-pCYC1	pCYC1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
pRef-pPGK1	pPGK1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
pRef-pTDH3	pTDH3-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p2a_empty	*URA3*, CEN/ARS	This study
p_UAS-cpTEF_1	UAS-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS-cpTEF_2	UAS-cpTEF_2-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS-cpTEF_3	UAS-cpTEF_3-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS-cpTEF_4	UAS-cpTEF_4-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS-cpTEF_5	UAS-cpTEF_5-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS-cpTEF_6	UAS-cpTEF_6-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS-cpTEF_7	UAS-cpTEF_7-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS-cpTEF_8	UAS-cpTEF_8-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS-cpTEF_9	UAS-cpTEF_9-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_1	cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_2	cpTEF_2-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_3	cpTEF_3-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_4	cpTEF_4-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_5	cpTEF_5-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_6	cpTEF_6-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_7	cpTEF_7-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_8	cpTEF_8-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_9	cpTEF_9-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_6-libA	cpTEF_6_libA-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_6-libB	cpTEF_6_libB-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_6-libC	cpTEF_6_libC-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_cpTEF_6-libD	cpTEF_6_libD-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
pU-TEF1-M	Carrier vector for multiple *TEF1* UAS integration, pJET backbone	This study
pU-TEF1-S	Carrier vector for single *TEF1* UAS integration, pJET backbone	This study
pU-CLB2-M	Carrier vector for multiple *CLB2* UAS integration, pJET backbone	This study
pU-CLB2-S	Carrier vector for single *CLB2* UAS integration, pJET backbone	This study
pU-CIT1-M	Carrier vector for multiple *CIT1* UAS integration, pJET backbone	This study
pU-CIT1-S	Carrier vector for single *CIT1* UAS integration, pJET backbone	This study
pDest	GG destination vector, AarI-SacB-AarI-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, ChlorR, CEN/ARS	This study
p_UAS$_{TEF1}$-1x	UAS$_{TEF1}$-cpTEF_1-TEF1_UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS$_{TEF1}$-2x	UAS$_{TEF1}$(2x)-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS$_{TEF1}$-3x	UAS$_{TEF1}$(3x)-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study

p_UAS$_{TEF1}$-4x	UAS$_{TEF1}$(4x)-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS$_{CLB2}$-1x	UAS$_{CLB2}$(1x)-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS$_{CLB2}$-2x	UAS$_{CLB2}$(2x)-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS$_{CLB2}$-3x	UAS$_{CLB2}$(3x)-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS$_{CIT1}$-1x	UAS$_{CIT1}$(1x)-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS$_{CIT1}$-2x	UAS$_{CIT1}$(2x)-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS$_{CIT1}$-3x	UAS$_{CIT1}$(3x)-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS$_{CIT1\text{-}CLB2}$	UAS$_{CIT1}$-UAS$_{CLB2}$-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study
p_UAS$_{CIT1\text{-}TEF1\text{-}CLB2}$	UAS$_{CIT1}$-UAS$_{TEF1}$-UAS$_{CLB2}$-cpTEF_1-5'UTR-yECitrine-tADH1, *URA3*, AmpR, CEN/ARS	This study

Table S.1.3: Primers used in this study for the construction of the truncated *TEF1* promoter library and the randomized minimal *TEF1* core promoter libraries. In all oligo and plasmid names, the *TEF1* promoter is shortly named as TEF and the core promoter as cpTEF.

Primer	Sequence (5' – 3')	Description
o_UAScpTEF_1	GAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAG GC**AATAAAAATTTTTATCACGTTT**	Construction primer for p_UAS-cpTEF_1
o_UAScpTEF_2	GAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAG GC**CTTTTTCTTGAAAATTTTTTTTTTTG**	Construction primer for p_UAS-cpTEF_2
o_UAScpTEF_3	GAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAG GC**GATTTTTTTCTCTTTCGATGAC**	Construction primer for p_UAS-cpTEF_3
o_UAScpTEF_4	GAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAG GC**GACCTCCCATTGATATTTAAG**	Construction primer for p_UAS-cpTEF_4
o_UAScpTEF_5	GAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAG GC**GTTAATAAACGGTCTTCAATTTC**	Construction primer for p_UAS-cpTEF_5
o_UAScpTEF_6	GAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAG GC**TCTCAAGTTTCAGTTTCATTTTTC**	Construction primer for p_UAS-cpTEF_6
o_UAScpTEF_7	GAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAG GC**CTTGTTCTATTACAACTTTTTTTAC**	Construction primer for p_UAS-cpTEF_7
o_UAScpTEF_8	GAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAG GC**ACTTCTTGCTCATTAGAAAGAAAG**	Construction primer for p_UAS-cpTEF_8
o_UAScpTEF_9	GAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAG GC**AAGCATAGCAATCTAATCTAAG**	Construction primer for p_UAS-cpTEF_9
o_cpTEF_1	CACGACGTTGTAAAACGACGGCCAGTGAATTC**AAT AAAAATTTTTATCACGTTTCTTTTTCTTG**	Construction primer for p_cpTEF_1
o_cpTEF_2	CACGACGTTGTAAAACGACGGCCAGTGAATTC**CTT TTTCTTGAAAATTTTTTTTTTTG**	Construction primer for p_cpTEF_2
o_cpTEF_3	CACGACGTTGTAAAACGACGGCCAGTGAATTC**GAT TTTTTTCTCTTTCGATGAC**	Construction primer for p_cpTEF_3
o_cpTEF_4	CACGACGTTGTAAAACGACGGCCAGTGAATTC**GAC CTCCCATTGATATTTAAG**	Construction primer for p_cpTEF_4
o_cpTEF_5	CACGACGTTGTAAAACGACGGCCAGTGAATTC**GTT AATAAACGGTCTTCAATTTC**	Construction primer for p_cpTEF_5
o_cpTEF_6	CACGACGTTGTAAAACGACGGCCAGTGAATTC**TCT CAAGTTTCAGTTTCATTTTTC**	Construction primer for p_cpTEF_6
o_cpTEF_7	CACGACGTTGTAAAACGACGGCCAGTGAATTC**CTT GTTCTATTACAACTTTTTTTAC**	Construction primer for p_cpTEF_7
o_cpTEF_8	CACGACGTTGTAAAACGACGGCCAGTGAATTC**ACT TCTTGCTCATTAGAAAGAAAG**	Construction primer for p_cpTEF_8
o_cpTEF_9	CACGACGTTGTAAAACGACGGCCAGTGAATTC**AAG CATAGCAATCTAATCTAAG**	Construction primer for p_cpTEF_9
o_BBsplit_fw	CGGCGTAATCATGGTCATAG	Forward primer to split p2a backbone
o_BBsplit_rv	CGGATAACAATTTCACACAGG	Reverse primer to split p2a backbone
o_BBUAScpTEF	GCCTTTTTCGACGAAGAAAAAGAAACGAGGCGGTC TC	Reverse primer for construction p_cpTEF
o_BBcpTEF	GAATTCACTGGCCGTCGTTTTAC	Reverse primer for construction p_UAS-cpTEF
o_DegeneracyA	AGTAAAAAAAGTTGTAATAGAACAAGAAAAANNN NNNNNNNNNNNNNNNGAATTCACTGGCCGTCGT TTTACAACGTC	Construction primer for p_cpTEF_6-libA

o_DegeneracyB	CTTTCTAATGAGCAAGAAGTAAAAAAAGTTNNNN NNNNNNNNNNNNNNTGAAACTGAAACTTGAGAG AATTCAC	Construction primer for p_cpTEF_6-libB
o_DegeneracyC	GATTAGATTGCTATGCTTTCTTTCTAATGAGNNNN NNNNNNNNNNNNNNGTAATAGAACAAGAAAAAT GAAAC	Construction primer for p_cpTEF_6-libC
o_DegeneracyD	CATTTTGTAATTAAAACTTAGATTAGATTGCTANN NNNNNNNNNNNNNNNNCAAGAAGTAAAAAAAGT TGTAATAG	Construction primer for p_cpTEF_6-libD

Table S.1.4: Sequences of the different core promoters obtained after truncation of the native 176 bp *TEF1* core promoter (cpTEF_1). The shortened sequence compared to the next cpTEF is indicated in bold.

Core promoter	Sequence (5' – 3')
cpTEF_1 (176 bp)	**AATAAAAATTTTTATCACGTTT**CTTTTTCTTGAAAATTTTTTTTTTTTGATTTTT TTCTCTTTCGATGACCTCCCATTGATATTTAAGTTAATAAACGGTCTTCAATTTC TCAAGTTTCAGTTTCATTTTTCTTGTTCTATTACAACTTTTTTTACTTCTTGCTCA TTAGAAAGAAA
cpTEF_2 (154 bp)	**CTTTTTCTTGAAAATTTTTTTTTTT**GATTTTTTTCTCTTTCGATGACCTCCCAT TGATATTTAAGTTAATAAACGGTCTTCAATTTCTCAAGTTTCAGTTTCATTTTTC TTGTTCTATTACAACTTTTTTTACTTCTTGCTCATTAGAAAGAAA
cpTEF_3 (129bp)	**GATTTTTTTCTCTTTCGAT**GACCTCCCATTGATATTTAAGTTAATAAACGGTCT TCAATTTCTCAAGTTTCAGTTTCATTTTTCTTGTTCTATTACAACTTTTTTTACTT CTTGCTCATTAGAAAGAAA
cpTEF_4 (110 bp)	**GACCTCCCATTGATATTTAA**GTTAATAAACGGTCTTCAATTTCTCAAGTTTCAG TTTCATTTTTCTTGTTCTATTACAACTTTTTTTACTTCTTGCTCATTAGAAAGAA A
cpTEF_5 (90 bp)	**GTTAATAAACGGTCTTCAATTT**CTCAAGTTTCAGTTTCATTTTTCTTGTTCTAT TACAACTTTTTTTACTTCTTGCTCATTAGAAAGAAA
cpTEF_6 (69 bp)	**TCTCAAGTTTCAGTTTCATTTTT**CTTGTTCTATTACAACTTTTTTTACTTCTTG CTCATTAGAAAGAAA
cpTEF_7 (46 bp)	**CTTGTTCTATTACAACTTTTTTT**ACTTCTTGCTCATTAGAAAGAAA
cpTEF_8 (23 bp)	**ACTTCTTGCTCATTAGAAAGA**AA
cpTEF_9 (2 bp)	AA

Table S.1.5: Upstream activating sequences (UAS) assembled in the pJET vector backbone (ThermoFisher). The underlined sequences represent the AarI restriction site, the actual UAS is indicated in bold. The abbreviations M and S are the designs for respectively multiple or single integration of the UASs in the destination vector pDest (Supplementary Table S.1.2).

UAS	Sequence (5' – 3')
TEF1-M	CACCTGCGAGTGGAG**ATAGCTTCAAAATGTTTCTACTCCTTTTTTACTCTTCCAGAT TTTCTCGGACTCCGCGCATCGCCGTACCACTTCAAAACACCCAAGCACAGCATACTA AATTTCCCCTCTTTCTTCCTCTAGGGTGTCGTTAATTACCCGTACTAAAGGTTTGG AAAAGAAAAAAGAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAGGC**GGAGCTCA GCAGGTG
TEF1-S	CACCTGCGAGTGGAG**ATAGCTTCAAAATGTTTCTACTCCTTTTTTACTCTTCCAGAT TTTCTCGGACTCCGCGCATCGCCGTACCACTTCAAAACACCCAAGCACAGCATACTA AATTTCCCCTCTTTCTTCCTCTAGGGTGTCGTTAATTACCCGTACTAAAGGTTTGG AAAAGAAAAAAGAGACCGCCTCGTTTCTTTTTCTTCGTCGAAAAAGGC**CGCTCTCA GCAGGTG
CIT1-M	CACCTGCGAGTGGAG**TAGAGATTACTACATATTCCAACAAGACCTTCGCAGGAAAG TATACCTAAACTAATTAAAGAAATCTCCGAAGTTCGCATTTCATTGAACGGCTCAA TTAATCTTTGTAAATATGAGCGTTTTTACGTTCACATTGCCTTTTTTTTTATGTA TTTACCTTGCATTTTTGTGCTAAAAGGCGTCACGTTTTTTTCCGCCGCAGCCGCCC GGAAATGAAAAGTATGACCCCCGCTAGACCAAAAATACTTTTGTGTTATTGGAGG ATCGCAATCCCT**GGAGCTCAGCAGGTG
CIT1-S	CACCTGCGAGTGGAG**TAGAGATTACTACATATTCCAACAAGACCTTCGCAGGAAAG TATACCTAAACTAATTAAAGAAATCTCCGAAGTTCGCATTTCATTGAACGGCTCAA TTAATCTTTGTAAATATGAGCGTTTTTACGTTCACATTGCCTTTTTTTTTATGTA TTTACCTTGCATTTTTGTGCTAAAAGGCGTCACGTTTTTTTCCGCCGCAGCCGCCC GGAAATGAAAAGTATGACCCCCGCTAGACCAAAAATACTTTTGTGTTATTGGAGG ATCGCAATCCCT**CGCTCTCAGCAGGTG
CLB2-M	CACCTGCGAGTGGAG**AGTGGAATTATTAGAATGACCACTACTCCTTCTAATCAAAC ACGCGGAAATAGCCGCCAAAAGACAGATTTTATTCCAAATGCGGGTAACTATTTGT ATAATATGTTTACATATTGAGCCCGTTTAGGAAAGTGCAAGTTCAAGGCACTAAT CAAAAAAGGAGATTTGTAAATATAGCGACCGAATCAGGAAAAGGTCAACAACGAA GTTCGCGATATGGATGAACTTCGGTGCCTGTCC**GGAGCTCAGCAGGTG
CLB2-S	CACCTGCGAGTGGAG**AGTGGAATTATTAGAATGACCACTACTCCTTCTAATCAAAC ACGCGGAAATAGCCGCCAAAAGACAGATTTTATTCCAAATGCGGGTAACTATTTGT ATAATATGTTTACATATTGAGCCCGTTTAGGAAAGTGCAAGTTCAAGGCACTAAT CAAAAAAGGAGATTTGTAAATATAGCGACCGAATCAGGAAAAGGTCAACAACGAA GTTCGCGATATGGATGAACTTCGGTGCCTGTCC**CGCTCTCAGCAGGTG

```
LOCUS       pRef-pTEF1       5940 bp    DNA     circular SYN 17-DEC-2017
DEFINITION  Join_product added to end of PCR_prod_p2a_111_VBB, overlap trimmed
ACCESSION   p2a111_EcoRI-TEF
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown
            Unclassified.
REFERENCE   1  (bases 1 to 5940)
  AUTHORS   Self
  JOURNAL   Unpublished.
COMMENT     SECID/File created by Clone Manager, Scientific & Educational Software
FEATURES             Location/Qualifiers
     misc_feature    complement(146..401)
                     /label=''lacZ
                     /SECDrawAs="Info only"
     misc_feature    402..604
                     /gene="UAS_TEF1"
                     /product="upstream activating sequence TEF1 gene"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    605..813
                     /gene="TEF1 core promoter"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     CDS             814..1530
                     /gene="'yECitrine"
                     /product="yeast enhanced yellow fluorescent protein"
                     /codon_start=3
                     /translation="V"
                     /SECDrawAs="Gene"
                     /SECStyleId=1
                     /SECName="yECitrine"
                     /SECDescr="yeast enhanced yellow fluorescent protein"
     misc_feature    1538..1740
                     /gene="ADH1t"
                     /product="ADH terminator"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    1809..1936
                     /gene="Cen6"
                     /product="yeast chromosome VI centromere sequence"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    1949..2322
                     /gene="ars4"
                     /product="ARS209; histone H4 autonomously replicating
                     sequence"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    2576..2807
                     /gene="pURA3"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     CDS             2808..3611
                     /gene="URA3"
                     /product="orotidine-5'-phosphate decarboxylase"
                     /SECDrawAs="Gene"
                     /SECStyleId=1
     misc_feature    3602..3679
                     /gene="tURA3"
                     /product="URA terminator"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    4061..4780
                     /gene="pUC ori"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     CDS             complement(4880..5740)
                     /gene="AmpR"
                     /SECDrawAs="Gene"
```

/SECStyleId=1
ORIGIN
1 tcgcgcgttt cggtgatgac ggtgaaaacc tctgacacat gcagctcccg gagacggtca
61 cagcttgtct gtaagcggat gccgggagca gacaagcccg tcagggcgcg tcagcgggtg
121 ttggcgggtg tcggggctgg cttaactatg cggcatcaga gcagattgta ctgagagtgc
181 accatatgcg gtgtgaaata ccgcacagat gcgtaaggag aaaataccgc atcaggcgcc
241 attcgccatt caggctgcgc aactgttggg aagggcgatc ggtgcgggcc tcttcgctat
301 tacgccagct ggcgaaaggg ggatgtgctg caaggcgatt aagttgggta acgccagggt
361 tttcccagtc acgacgttgt aaaacgacgg ccagtgaatt catagcttca aaatgtttct
421 actccttttt tactcttcca gattttctcg gactccgcgc atcgccgtac cacttcaaaa
481 cacccaagca cagcatacta aatttcccct ctttcttcct ctagggtgtc gttaattacc
541 cgtactaaag gtttggaaaa gaaaaaagag accgcctcgt ttctttttct tcgtcgaaaa
601 aggcaataaa aatttttatc acgtttcttt ttcttgaaaa ttttttttt tgatttttt
661 ctctttcgat gacctcccat tgatatttaa gttaataaac ggtcttcaat ttctcaagtt
721 tcagtttcat ttttcttgtt ctattacaac tttttttact tcttgctcat tagaaagaaa
781 gcatagcaat ctaatctaag ttttaattac aaaatgtcta aaggtgaaga attattcact
841 ggtgttgtcc caattttggt tgaattagat ggtgatgtta atggtcacaa attttctgtc
901 tccggtgaag gtgaaggtga tgctacttac ggtaaattga ccttaaaatt tatttgtact
961 actggtaaat tgccagttcc atggccaacc ttagtcacta ctttaggtta tggtttgatg
1021 tgttttgcta gatacccaga tcatatgaaa caacatgact ttttcaagtc tgccatgcca
1081 gaaggttatg ttcaagaaag aactattttt ttcaaagatg acggtaacta caagaccaga
1141 gctgaagtca agtttgaagg tgatacctta gttaatagaa tcgaattaaa aggtattgat
1201 tttaaagaag atggtaacat tttaggtcac aaattggaat acaactataa ctctcacaat
1261 gtttacatca tggctgacaa acaaaagaat ggtatcaaag ttaacttcaa aattagacac
1321 aacattgaag atggttctgt tcaattagct gaccattatc aacaaaatac tccaattggt
1381 gatggtccag tcttgttacc agacaaccat tacttatcct atcaatctgc cttatccaaa
1441 gatccaaacg aaaagagaga ccacatggtc ttgttagaat ttgttactgc tgctggtatt
1501 acccatggta tggatgaatt gtacaaataa ggcgcgccac ttctaaataa gcgaatttct
1561 tatgatttat gatttttatt attaaataag ttataaaaaa aataagtgta tacaaatttt
1621 aaagtgactc ttaggtttta aaacgaaaat tcttattctt gagtaactct ttcctgtagg
1681 tcaggttgct ttctcaggta tagtatgagg tcgctcttat tgaccacacc tctaccggca
1741 cccggggagc gtcccaaaac cttctcaagc aaggttttca gtataatgtt acatgcgtac
1801 acccgtcggg tccttttcat cacgtgctat aaaaataatt ataatttaaa ttttttaata
1861 taaatatata aattaaaaat agaaagtaaa aaaagaaatt aaagaaaaaa tagtttttgt
1921 tttccgaaga tgtaaaagac tctagggggа tcgccaacaa atactacctt ttatcttgct
1981 cttcctgctc tcaggtatta atgccgaatt gtttcatctt gtctgtgtag aagaccacac
2041 acgaaaatcc tgtgatttta cattttactt atcgttaatc gaatgtatat ctatttaatc
2101 tgcttttctt gtctaataaa tatatatgta aagtacgctt tttgttgaaa ttttttaaac
2161 ctttgtttat ttttttttct tcattccgta actcttctac cttctttatt tactttctaa
2221 aatccaaata caaaacataa aaataaataa acacagagta aattcccaaa ttattccatc
2281 attaaaagat acgaggcgcg tgtaagttac aggcaagcga tccgtcctaa gaaaccatta
2341 ttatcatgac attaacctat aaaaatаggc gtatcacgag gccctttcgt ctcgcgcgtt
2401 tcggtgatga cggtgaaaac ctctgacaca tgcagctccc ggagacggtc acagcttgtc
2461 tgtaagcgga tgccgggagc agacaagccc gtcagggcgc gtcagcgggt gttggcgggt
2521 gtcggggctg gcttaactat gcggcatcag agcagattgt actgagagtg caccatacca
2581 cagcttttca attcaattca tcatttttt tttattcttt tttttgattt cggtttcttt
2641 gaaatttttt tgattcggta atctccgaac agaaggaaga acgaaggaag gagcacagac
2701 ttagattggt atatatacgc atatgtagtg ttgaagaaac atgaaattgc ccagtattct
2761 taacccaact gcacagaaca aaaacctgca ggaaacgaag ataaatcatg tcgaaagcta
2821 catataagga acgtgctgct actcatccta gtcctgttgc tgccaagcta tttaatatca
2881 tgcacgaaaa gcaaacaaac ttgtgtgctt cattggatgt tcgtaccacc aaggaattac
2941 tggagttagt tgaagcatta ggtcccaaaa tttgtttact aaaaacacat gtggatatct
3001 tgactgattt ttccatggag ggcacagtta agccgctaaa ggcattatcc gccaagtaca
3061 atttttact cttcgaagac agaaaatttg ctgacattgg taatacagtc aaattgcagt
3121 actctgcggg tgtatacaga atagcagaat gggcagacat tacgaatgca cacggtgtgg
3181 tgggcccagg tattgttagc ggtttgaagc aggcggcaga agaagtaaca aaggaaccta
3241 gaggcctttt gatgttagca gaattgtcat gcaagggctc cctatctact ggagaatata
3301 ctaagggtac tgttgacatt gcgaagagcg acaaagattt tgttatcggc tttattgctc
3361 aaagagacat gggtggaaga gatgaaggtt acgattggtt gattatgaca cccggtgtgg
3421 gtttagatga caagggagac gcattgggtc aacagtatag aaccgtggat gatgtggtct
3481 ctacaggatc tgacattatt attgttggaa gaggactatt tgcaaaggga agggatgcta
3541 aggtagaggg tgaacgttac agaaaagcag gctgggaagc atatttgaga agatgcggcc
3601 agcaaaacta aaaaactgta ttataagtaa atgcatgtat actaaactca caaattagag
3661 cttcaattta attatatcag ttattaccct atgcggtgtg aaatacggcg taatcatggt
3721 catagctgtt tcctgtgtga aattgttatc cgctcacaat tccacacaac atacgagccg
3781 gaagcataaa gtgtaaagcc tggggtgcct aatgagtgag ctaactcaca ttaattgcgt
3841 tgcgctcact gcccgctttc cagtcgggaa acctgtcgtg ccagctgcat taatgaatcg
3901 gccaacgcgc ggggagaggc ggtttgcgta ttgggcgctc ttccgcttcc tcgctcactg
3961 actcgctgcg ctcggtcgtt cggctgcggc gagcggtatc agctcactca aaggcggtaa
4021 tacggttatc cacagaatca ggggataacg caggaaagaa catgtgagca aaaggccagc

```
4081 aaaaggccag gaaccgtaaa aaggccgcgt tgctggcgtt tttccatagg ctccgccccc
4141 ctgacgagca tcacaaaaat cgacgctcaa gtcagaggtg gcgaaacccg acaggactat
4201 aaagatacca ggcgtttccc cctggaagct ccctcgtgcg ctctcctgtt ccgaccctgc
4261 cgcttaccgg atacctgtcc gcctttctcc cttcgggaag cgtggcgctt tctcatagct
4321 cacgctgtag gtatctcagt tcggtgtagg tcgttcgctc caagctgggc tgtgtgcacg
4381 aacccccgt tcagcccgac cgctgcgcct tatccggtaa ctatcgtctt gagtccaacc
4441 cggtaagaca cgacttatcg ccactggcag cagccactgg taacaggatt agcagagcga
4501 ggtatgtagg cggtgctaca gagttcttga agtggtggcc taactacggc tacactagaa
4561 gaacagtatt tggtatctgc gctctgctga agccagttac cttcggaaaa agagttggta
4621 gctcttgatc cggcaaacaa accaccgctg gtagcggtgg ttttttttgtt tgcaagcagc
4681 agattacgcg cagaaaaaaa ggatctcaag aagatccttt gatcttttct acggggtctg
4741 acgctcagtg gaacgaaaac tcacgttaag ggattttggt catgagatta tcaaaaagga
4801 tcttcaccta gatcctttta aattaaaaat gaagttttaa atcaatctaa agtatatatg
4861 agtaaacttg gtctgacagt taccaatgct taatcagtga ggcacctatc tcagcgatct
4921 gtctatttcg ttcatccata gttgcctgac tccccgtcgt gtagataact acgatacggg
4981 agggcttacc atctggcccc agtgctgcaa tgataccgcg agacccacgc tcaccggctc
5041 cagatttatc agcaataaac cagccagccg gaagggccga gcgcagaagt ggtcctgcaa
5101 ctttatccgc ctccatccag tctattaatt gttgccggga agctagagta agtagttcgc
5161 cagttaatag tttgcgcaac gttgttgcca ttgctacagg catcgtggtg tcacgctcgt
5221 cgtttggtat ggcttcattc agctccggtt cccaacgatc aaggcgagtt acatgatccc
5281 ccatgttgtg caaaaaagcg gttagctcct tcggtcctcc gatcgttgtc agaagtaagt
5341 tggccgcagt gttatcactc atggttatgg cagcactgca taattctctt actgtcatgc
5401 catccgtaag atgcttttct gtgactggtg agtactcaac caagtcattc tgagaatagt
5461 gtatgcggcg accgagttgc tcttgcccgg cgtcaatacg ggataatacc gcgccacata
5521 gcagaacttt aaaagtgctc atcattggaa aacgttcttc ggggcgaaaa ctctcaagga
5581 tcttaccgct gttgagatcc agttcgatgt aacccactcg tgcacccaac tgatcttcag
5641 catcttttac tttcaccagc gtttctgggt gagcaaaaac aggaaggcaa aatgccgcaa
5701 aaaagggaat aagggcgaca cggaaatgtt gaatactcat actcttcctt tttcaatatt
5761 attgaagcat ttatcagggt tattgtctca tgagcggata catatttgaa tgtatttaga
5821 aaaataaaca aatagggttt ccgcgcacat ttccccgaaa agtgccacct gacgtctaag
5881 aaaccattat tatcatgaca ttaacctata aaaataggcg tatcacgagg ccctttcgtc
//
```

Figure S.1.1: Annotated Genbank file of the UAS_{TEF1}-cpTEF_1-5'UTR_{TEF1}-yECitrine-tADH1 transcription unit in pRef-pTEF1 and p_UAS-cpTEF_1. The *TEF1* 5'UTR is indicated in bold and underlined. The UAS_{TEF1} is underlined and the *TEF1* core promoter is indicated in bold. The respective sequences of the truncated *TEF1* core promoter are represented in Supplementary Table S.1.4. Primer sites are indicated in yellow and are respectively o_BBcpTEF, o_BBUAScpTEF, o_BBsplit_fw and o_BBsplit_rv in order of occurrence. For the p_cpTEF plasmids, the UAS_{TEF1} sequence in front of the core promoter was not present.

```
LOCUS       AarI-SacB-AarI-c        3048 bp    DNA     linear   SYN 13-NOV-2017
DEFINITION  p2a_chlorR_EcoRI-AarI-SacB-AarI-TEF1core-yECit-tADH1-XmaI cut 5668
            to 1151
ACCESSION   AarI-SacB-AarI-c
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown
            Unclassified.
REFERENCE   1  (bases 1 to 3048)
  AUTHORS   Self
  JOURNAL   Unpublished.
COMMENT     SECID/File created by Clone Manager, Scientific & Educational Software
FEATURES             Location/Qualifiers
     misc_signal     9..15
                     /label=AarI
                     /SECDrawAs="Label"
     misc_feature    73..173
                     /gene="promoter 14'''"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     CDS             complement(214..279)
                     /gene="dyad symmetry"
                     /SECDrawAs="Gene"
                     /SECStyleId=1
     CDS             373..1794
                     /gene="sacB"
                     /product="levansucrase precursor"
                     /SECDrawAs="Gene"
                     /SECStyleId=1
     misc_signal     1898..1904
                     /label=AarI
                     /SECDrawAs="Label"
     misc_feature    1913..2121
                     /gene="TEF1 core promoter"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     CDS             2122..2838
                     /gene="'yECitrine"
                     /product="yeast enhanced yellow fluorescent protein"
                     /codon_start=3
                     /translation="V"
                     /SECDrawAs="Gene"
                     /SECStyleId=1
                     /SECName="yECitrine"
                     /SECDescr="yeast enhanced yellow fluorescent protein"
     misc_feature    2846..3048
                     /gene="tADH1"
                     /product="ADH1 terminator"
                     /SECDrawAs="Region"
                     /SECStyleId=1
ORIGIN
        1 ggaggagtgc aggtgccgct tacagacaag ctgtgaccgt ctccgggaga gctcgatatc
       61 ccgggcggcc gccttcattc tataagtttc ttgacatctt ggccggcata tggtataata
      121 gggaaatttc catggcggcc gctctagaag aagcttggga tccgtcgacc tcgaattggt
      181 aaatcgcgcg ggtttgttac tgataaagca ggcaagacct aaaatgtgta aagggcaaag
      241 tgtatacttt ggcgtcaccc cttacatatt ttaggtcttt ttttattgtg cgtaactaac
      301 ttgccatctt caaacaggag ggctggaaga agcagaccgc taacacagta cataaaaaag
      361 gagacatgaa cgatgaacat caaaaagttt gcaaaacaag caacagtatt aacctttact
      421 accgcactgc tggcaggagg cgcaactcaa gcgtttgcga aagaaacgaa ccaaaagcca
      481 tataaggaaa catacggcat ttcccatatt acacgccatg atatgctgca aatccctgaa
      541 cagcaaaaaa atgaaaaata tcaagttcct gaattcgatt cgtccacaat taaaaatatc
      601 tcttctgcaa aaggcctgga cgtttgggac agctggccat tacaaaacgc tgacggcact
      661 gtcgcaaact atcacggcta ccacatcgtc tttgcattag ccggagatcc taaaaatgcg
      721 gatgacacat cgatttacat gttctatcaa aaagtcggcg aaacttctat tgacagctgg
      781 aaaaacgctg gccgcgtctt taaagacagc gacaaattcg atgcaaatga ttctatccta
      841 aaagaccaaa cacaagaatg gtcaggttca gccacattta catctgacgg aaaaatccgt
      901 ttattctaca ctgatttctc cggtaaacat tacggcaaac aaacactgac aactgcacaa
      961 gttaacgtat cagcatcaga cagctctttg aacatcaacg gtgtagagga ttataaatca
     1021 atctttgacg gtgacggaaa aacgtatcaa aatgtacagc agttcatcga tgaaggcaac
     1081 tacagctcag gcgacaacca tacgctgaga gatcctcact acgtagaaga taaaggccac
```

```
1141 aaatacttag tatttgaagc aaacactgga actgaagatg gctaccaagg cgaagaatct
1201 ttatttaaca aagcatacta tggcaaaagc acatcattct tccgtcaaga aagtcaaaaa
1261 cttctgcaaa gcgataaaaa acgcacggct gagttagcaa acggcgctct cggtatgatt
1321 gagctaaacg atgattacac actgaaaaaa gtgatgaaac cgctgattgc atctaacaca
1381 gtaacagatg aaattgaacg cgcgaacgtc tttaaaatga acggcaaatg gtacctgttc
1441 actgactccc gcggatcaaa aatgacgatt gacggcatta cgtctaacga tatttacatg
1501 cttggttatg tttctaattc tttaactggc ccatacaagc cgctgaacaa aactggcctt
1561 gtgttaaaaa tggatcttga tcctaacgat gtaaccttta cttactcaca cttcgctgta
1621 cctcaagcga aaggaaacaa tgtcgtgatt acaagctata tgacaaacag aggattctac
1681 gcagacaaac aatcaacgtt tgcgccaagc ttcctgctga acatcaaagg caagaaaaca
1741 tctgttgtca aagacagcat ccttgaacaa ggacaattaa cagttaacaa ataaaaacgc
1801 aaaagaaaat gccgatatcc tattggcatt ttcttttatt tcttatcaac ataaaggtga
1861 atcccatagg gcaggagcta aggaagctaa aatggagcac ctgcctcacg ct
1921 tttttatcac gtttcttttt cttgaaaatt ttttttttg attttttct ctttcgatga
1981 cctcccattg atatttaagt taataaacgg tcttcaattt ctcaagtttc agtttcattt
2041 ttcttgttct attacaactt tttttacttc ttgctcatta gaaagaaagc atagcaatct
2101 aatctaagtt ttaattacaa aatgtctaaa ggtgaagaat tattcactgg tgttgtccca
2161 attttggttg aattagatgg tgatgttaat ggtcacaaat tttctgtctc cggtgaaggt
2221 gaaggtgatg ctacttacgg taaattgacc ttaaaattta tttgtactac tggtaaattg
2281 ccagttccat ggccaacctt agtcactact ttaggttatg gtttgatgtg ttttgctaga
2341 tacccagatc atatgaaaca acatgacttt ttcaagtctg ccatgccaga aggttatgtt
2401 caagaaagaa ctattttttt caaagatgac ggtaactaca agaccagagc tgaagtcaag
2461 tttgaaggtg ataccttagt taatagaatc gaattaaaag gtattgattt taaagaagat
2521 ggtaacattt taggtcacaa attggaatac aactataact ctcacaatgt ttacatcatg
2581 gctgacaaac aaaagaatgg tatcaaagtt aacttcaaaa ttagacacaa cattgaagat
2641 ggttctgttc aattagctga ccattatcaa caaaatactc caattggtga tggtccagtc
2701 ttgttaccag acaaccatta cttatcctat caatctgcct tatccaaaga tccaaacgaa
2761 aagagagacc acatggtctt gttagaattt gttactgctg ctggtattac ccatggtatg
2821 gatgaattgt acaaataagg cgcgccactt ctaaataagc gaatttctta tgatttatga
2881 tttttattat taaataagtt ataaaaaaaa taagtgtata caaattttaa agtgactctt
2941 aggttttaaa acgaaaattc ttattcttga gtaactcttt cctgtaggtc aggttgcttt
3001 ctcaggtata gtatgaggtc gctcttattg accacacctc taccggca
//
```

Figure S.1.2: Annotated Genbank file of the AarI-SacB-AarI-cpTEF_1-5'UTR$_{TEF1}$-yECitrine-tADH1 transcription unit of the destination plasmid for UAS library construction using yUGG. The *TEF1* 5'UTR is indicated in bold and underlined. The cpTEF_1 sequence is indicated in bold.

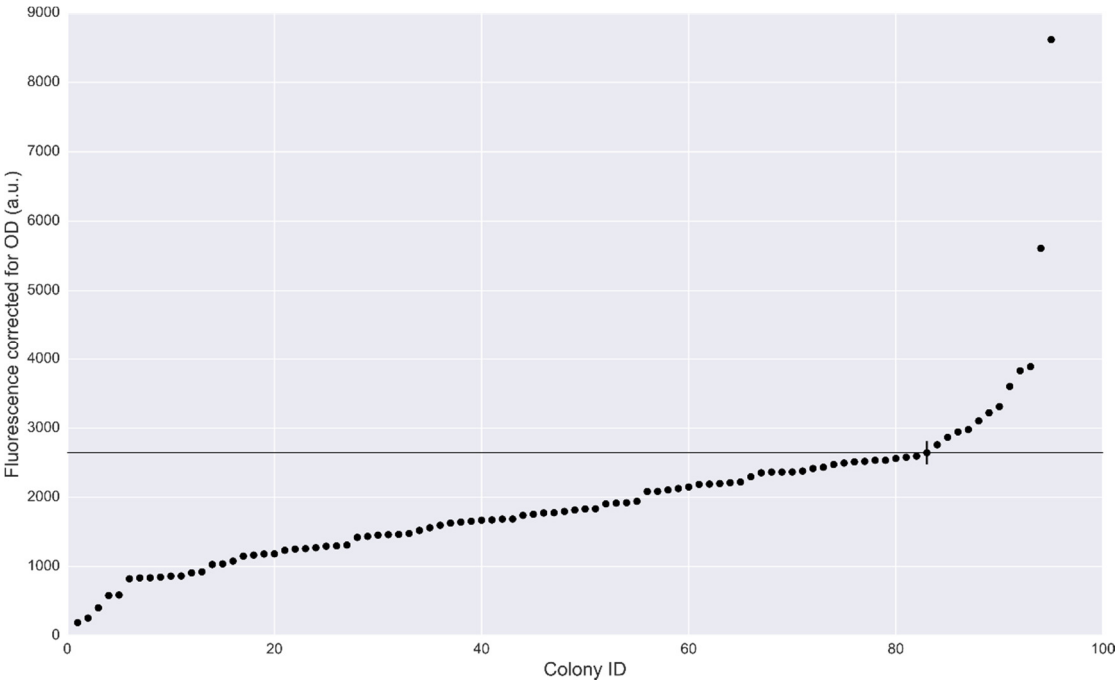

Figure S.1.3: Scatter plot of random core promoter library cpTEF_6-libA. The horizontal line represents the mean fluorescence corrected for OD of the native cpTEF_6 which was grown as biological triplicate. Error bars representing the standard error are a consequence of OD correction with biological triplicates of sRef-bl and the medium.

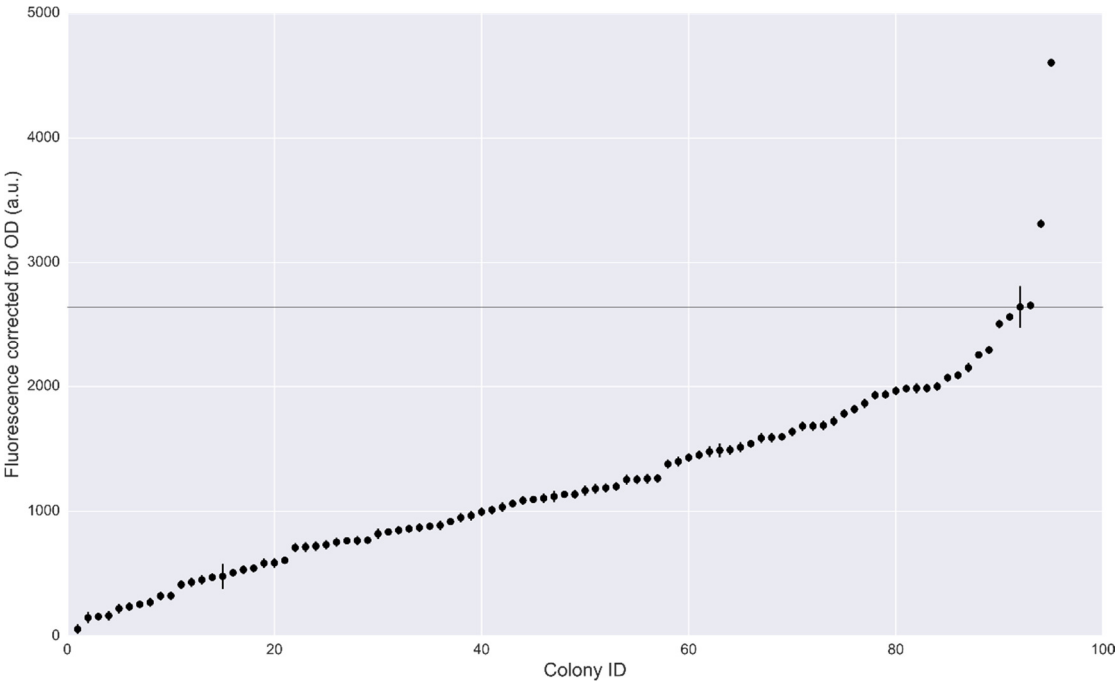

Figure S.1.4: Scatter plot of random core promoter library cpTEF_6-libB. The horizontal line represents the mean fluorescence corrected for OD of the native cpTEF_6 which was grown as biological triplicate. Error bars representing the standard error are a consequence of OD correction with biological triplicates of sRef-bl and the medium.

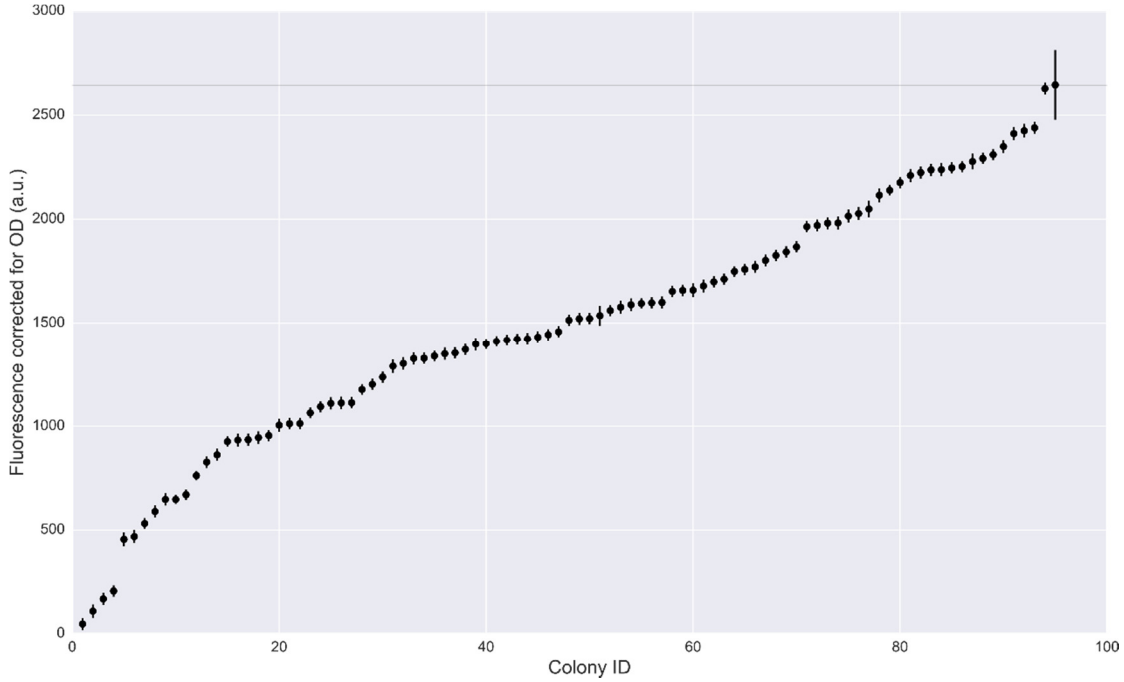

Figure S.1.5: Scatter plot of random core promoter library cpTEF_6-libC. The horizontal line represents the mean fluorescence corrected for OD of the native cpTEF_6 which was grown as biological triplicate. Error bars representing the standard error are a consequence of OD correction with biological triplicates of sRef-bl and the medium.

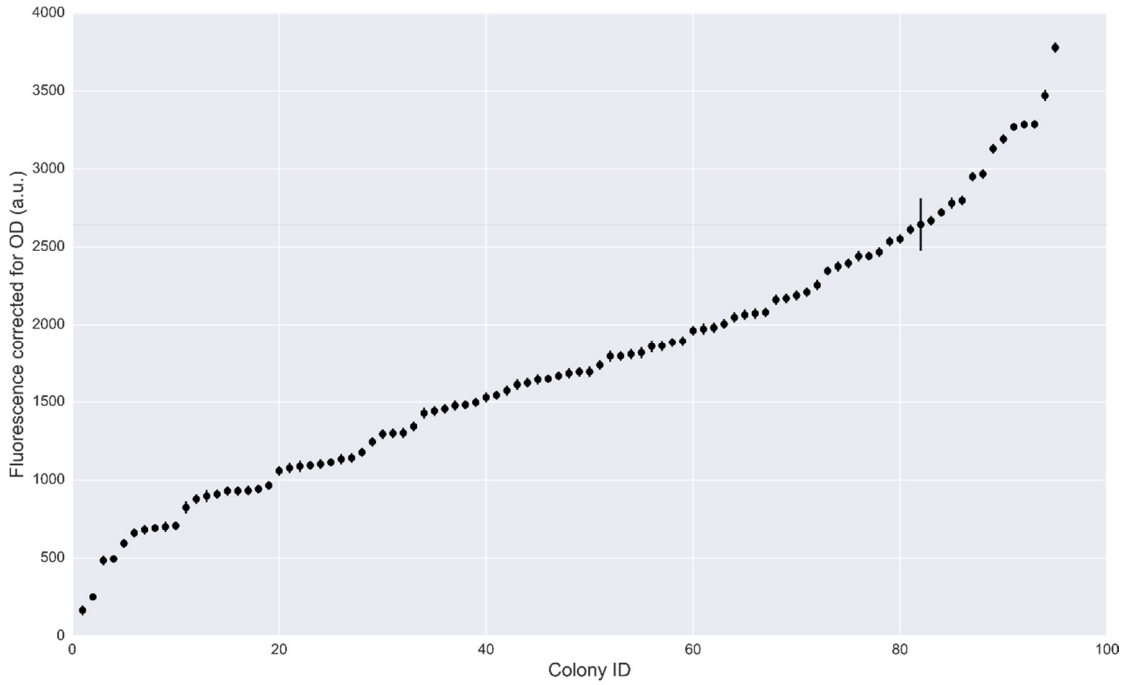

Figure S.1.6: Scatter plot of random core promoter library cpTEF_6-libD. The horizontal line represents the mean fluorescence corrected for OD of the native cpTEF_6 which was grown as biological triplicate. Error bars representing the standard error are a consequence of OD correction with biological triplicates of sRef-bl and the medium.

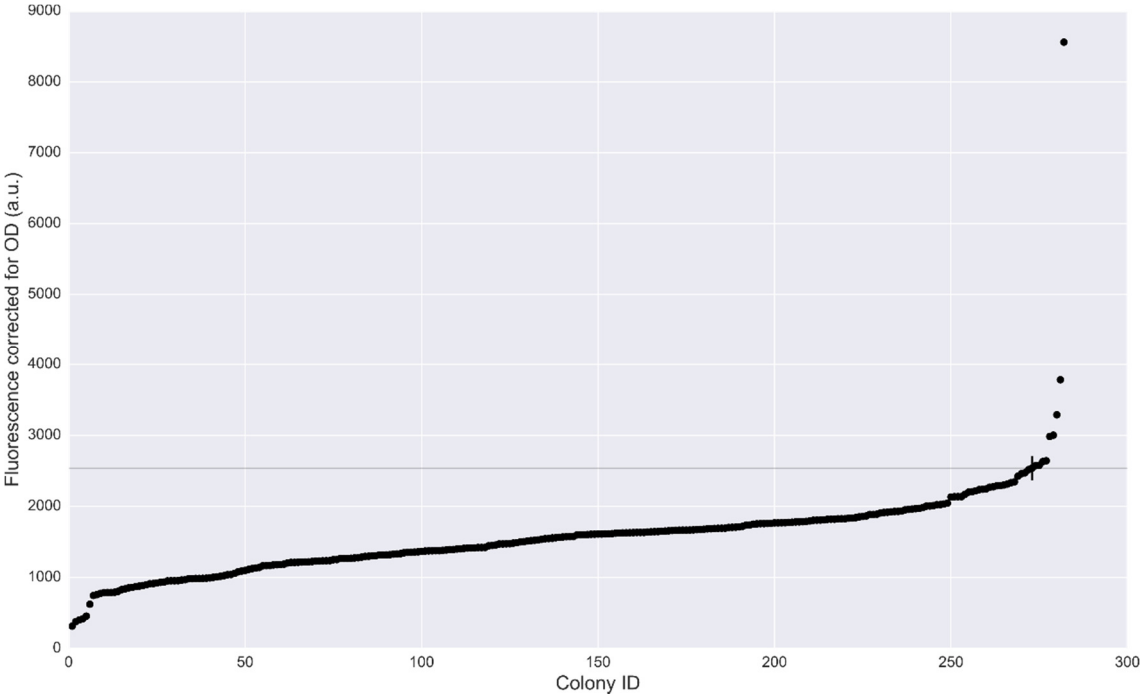

Figure S.1.7: Scatter plot of random core promoter library cpTEF_6-libA for 281 randomly chosen colonies. The horizontal line represents the mean fluorescence corrected for OD of the native cpTEF_6 which was grown as biological triplicate. Error bars representing the standard error are a consequence of OD correction with biological triplicates of sRef-bl and the medium.

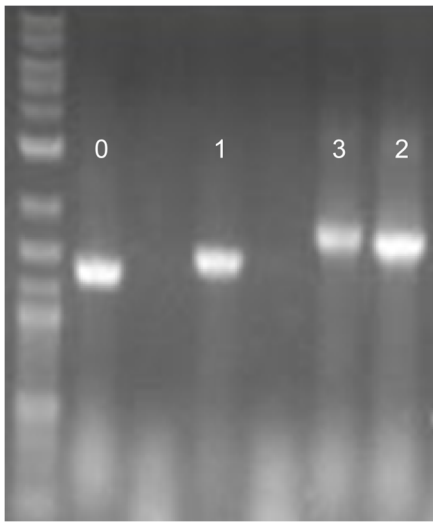

Figure S.1.8: Colony PCR example showing the distinction between zero, one, two and three incorporated upstream activated sequences (UAS) in the yUGG destination vector. The reference marker is the 2-log DNA ladder (New England Biolabs).

S.2 APPENDIX CHAPTER 4

Table S.2.1: Overview of all strains and plasmids used in this study. The predicted values of the protein abundances (PPA) are generated by the PLS regression model. PPA values that were predicted via reverse engineering are indicated with an asterisk. The plasmid backbone is a low copy backbone with a *CEN/ARS4* ori, the *URA3* auxotrophic marker and the *TEF2p-mCherry-PGK1t* transcription unit. All yeast strains were derived from the S288c laboratory strain BY4742.

Strain	Genotype	PPA	Plasmid
BY4742	*Matα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*	-	-
sTemplate1	RPL8Ap-nativeRPL8A_UTR-yECitrine-ADH1t	5.50*	pTemplate1
sTemplate2	TEF1coreP-nativeTEF1_UTR-yECitrine-ADH1t	6.42*	pTemplate2
sTemplate3	RPL8Ap-nativeRPL8A_UTR-mTFP1-ADH1t	6.24*	pTemplate3
sTemplate4	TEF1p-nativeTEF1_UTR-*RcTal1*-ADH1t	6.84*	pTemplate4
s_yC^{I}-1	RPL8Ap-UTRa1-yECitrine-ADH1t	5.46	p_yC^{I}-1
s_yC^{I}-2	RPL8Ap-UTRa2-yECitrine-ADH1t	3.38	p_yC^{I}-2
s_yC^{I}-3	RPL8Ap-UTRa3-yECitrine-ADH1t	6.64	p_yC^{I}-3
s_yC^{I}-4	RPL8Ap-UTRa4-yECitrine-ADH1t	4.05	p_yC^{I}-4
s_yC^{I}-5	RPL8Ap-UTRa5-yECitrine-ADH1t	4.96	p_yC^{I}-5
s_yC^{I}-6	RPL8Ap-UTRa6-yECitrine-ADH1t	2.67	p_yC^{I}-6
s_yC^{I}-7	RPL8Ap-UTRa7-yECitrine-ADH1t	5.83	p_yC^{I}-7
s_yC^{I}-8	RPL8Ap-UTRa8-yECitrine-ADH1t	3.08	p_yC^{I}-8
s_yC^{I}-9	RPL8Ap-UTRa9-yECitrine-ADH1t	5.82	p_yC^{I}-9
s_yC^{I}-10	RPL8Ap-UTRa10-yECitrine-ADH1t	6.62	p_yC^{I}-10
s_yC^{I}-11	RPL8Ap-UTRa11-yECitrine-ADH1t	5.12	p_yC^{I}-11
s_yC^{I}-12	RPL8Ap-UTRa12-yECitrine-ADH1t	5.72	p_yC^{I}-12
s_yC^{I}-13	RPL8Ap-UTRa13-yECitrine-ADH1t	4.11	p_yC^{I}-13
s_yC^{I}-14	RPL8Ap-UTRa14-yECitrine-ADH1t	4.64	p_yC^{I}-14
s_yC^{I}-15	RPL8Ap-UTRa15-yECitrine-ADH1t	2.69	p_yC^{I}-15
s_yC^{I}-16	RPL8Ap-UTRa16-yECitrine-ADH1t	3.40	p_yC^{I}-16
s_yC^{II}-1	TEF1coreP-UTRa1-yECitrine-ADH1t	5.46	p_yC^{II}-1
s_yC^{II}-2	TEF1coreP-UTRa2-yECitrine-ADH1t	3.38	p_yC^{II}-2
s_yC^{II}-3	TEF1coreP-UTRa3-yECitrine-ADH1t	6.64	p_yC^{II}-3
s_yC^{II}-4	TEF1coreP-UTRa4-yECitrine-ADH1t	4.05	p_yC^{II}-4
s_yC^{II}-5	TEF1coreP-UTRa5-yECitrine-ADH1t	4.96	p_yC^{II}-5
s_yC^{II}-6	TEF1coreP-UTRa6-yECitrine-ADH1t	2.67	p_yC^{II}-6
s_yC^{II}-7	TEF1coreP-UTRa7-yECitrine-ADH1t	5.83	p_yC^{II}-7
s_yC^{II}-8	TEF1coreP-UTRa8-yECitrine-ADH1t	3.08	p_yC^{II}-8
s_yC^{II}-9	TEF1coreP-UTRa9-yECitrine-ADH1t	5.82	p_yC^{II}-9
s_yC^{II}-10	TEF1coreP-UTRa10-yECitrine-ADH1t	6.62	p_yC^{II}-10
s_yC^{II}-11	TEF1coreP-UTRa11-yECitrine-ADH1t	5.12	p_yC^{II}-11
s_yC^{II}-12	TEF1coreP-UTRa12-yECitrine-ADH1t	5.72	p_yC^{II}-12
s_yC^{II}-13	TEF1coreP-UTRa13-yECitrine-ADH1t	4.11	p_yC^{II}-13
s_yC^{II}-14	TEF1coreP-UTRa14-yECitrine-ADH1t	4.64	p_yC^{II}-14
s_yC^{II}-15	TEF1coreP-UTRa15-yECitrine-ADH1t	2.69	p_yC^{II}-15
s_yC^{II}-16	TEF1coreP-UTRa16-yECitrine-ADH1t	3.40	p_yC^{II}-16
s_yC^{III}-1	RPL8Ap-UTRb1-mTFP1-ADH1t	7.25	p_yC^{III}-1
s_yC^{III}-2	RPL8Ap-UTRb2-mTFP1-ADH1t	4.57	p_yC^{III}-2
s_yC^{III}-3	RPL8Ap-UTRb3-mTFP1-ADH1t	5.88	p_yC^{III}-3

s_yC^{III}-4	RPL8Ap-UTRb4-mTFP1-ADH1t	3.21	p_yC^{III}-4
s_yC^{III}-5	RPL8Ap-UTRb5-mTFP1-ADH1t	6.74	p_yC^{III}-5
s_yC^{III}-6	RPL8Ap-UTRb6-mTFP1-ADH1t	4.11	p_yC^{III}-6
s_yC^{III}-7	RPL8Ap-UTRb7-mTFP1-ADH1t	5.38	p_yC^{III}-7
s_yC^{III}-8	RPL8Ap-UTRb8-mTFP1-ADH1t	2.74	p_yC^{III}-8
s_yC^{III}-9	RPL8Ap-UTRb9-mTFP1-ADH1t	7.26	p_yC^{III}-9
s_yC^{III}-10	RPL8Ap-UTRb10-mTFP1-ADH1t	4.57	p_yC^{III}-10
s_yC^{III}-11	RPL8Ap-UTRb11-mTFP1-ADH1t	6.69	p_yC^{III}-11
s_yC^{III}-12	RPL8Ap-UTRb12-mTFP1-ADH1t	3.97	p_yC^{III}-12
s_yC^{III}-13	RPL8Ap-UTRb13-mTFP1-ADH1t	6.14	p_yC^{III}-13
s_yC^{III}-14	RPL8Ap-UTRb14-mTFP1-ADH1t	3.42	p_yC^{III}-14
s_yC^{III}-15	RPL8Ap-UTRb15-mTFP1-ADH1t	5.67	p_yC^{III}-15
s_yC^{III}-16	RPL8Ap-UTRb16-mTFP1-ADH1t	2.75	p_yC^{III}-16
s_yC^{IV}-1	TEF1coreP-UTRc1-yECitrine-ADH1t	5.58	p_yC^{IV}-1
s_yC^{IV}-2	TEF1coreP-UTRc2-yECitrine-ADH1t	6.55	p_yC^{IV}-2
s_yC^{IV}-3	TEF1coreP-UTRc3-yECitrine-ADH1t	3.89	p_yC^{IV}-3
s_yC^{IV}-4	TEF1coreP-UTRc4-yECitrine-ADH1t	4.63	p_yC^{IV}-4
s_yC^{IV}-5	TEF1coreP-UTRc5-yECitrine-ADH1t	4.93	p_yC^{IV}-5
s_yC^{IV}-6	TEF1coreP-UTRc6-yECitrine-ADH1t	5.78	p_yC^{IV}-6
s_yC^{IV}-7	TEF1coreP-UTRc7-yECitrine-ADH1t	2.56	p_yC^{IV}-7
s_yC^{IV}-8	TEF1coreP-UTRc8-yECitrine-ADH1t	3.16	p_yC^{IV}-8
s_yC^{IV}-9	TEF1coreP-UTRc9-yECitrine-ADH1t	5.43	p_yC^{IV}-9
s_yC^{IV}-10	TEF1coreP-UTRc10-yECitrine-ADH1t	6.51	p_yC^{IV}-10
s_yC^{IV}-11	TEF1coreP-UTRc11-yECitrine-ADH1t	3.36	p_yC^{IV}-11
s_yC^{IV}-12	TEF1coreP-UTRc12-yECitrine-ADH1t	4.52	p_yC^{IV}-12
s_yC^{IV}-13	TEF1coreP-UTRc13-yECitrine-ADH1t	4.78	p_yC^{IV}-13
s_yC^{IV}-14	TEF1coreP-UTRc14-yECitrine-ADH1t	5.92	p_yC^{IV}-14
s_yC^{IV}-15	TEF1coreP-UTRc15-yECitrine-ADH1t	2.75	p_yC^{IV}-15
s_yC^{IV}-16	TEF1coreP-UTRc16-yECitrine-ADH1t	4.02	p_yC^{IV}-16
s_yC^{V}-1	RPL8Ap-UTRa1-mTFP1-ADH1t	5.87*	p_yC^{V}-1
s_yC^{V}-2	RPL8Ap-UTRa2-mTFP1-ADH1t	3.58*	p_yC^{V}-2
s_yC^{V}-3	RPL8Ap-UTRa3-mTFP1-ADH1t	6.91*	p_yC^{V}-3
s_yC^{V}-4	RPL8Ap-UTRa4-mTFP1-ADH1t	4.28*	p_yC^{V}-4
s_yC^{V}-5	RPL8Ap-UTRa5-mTFP1-ADH1t	5.13*	p_yC^{V}-5
s_yC^{V}-6	RPL8Ap-UTRa6-mTFP1-ADH1t	2.43*	p_yC^{V}-6
s_yC^{V}-7	RPL8Ap-UTRa7-mTFP1-ADH1t	5.97*	p_yC^{V}-7
s_yC^{V}-8	RPL8Ap-UTRa8-mTFP1-ADH1t	3.08*	p_yC^{V}-8
s_yC^{V}-9	RPL8Ap-UTRa9-mTFP1-ADH1t	7.02*	p_yC^{V}-9
s_yC^{V}-10	RPL8Ap-UTRa10-mTFP1-ADH1t	7.05*	p_yC^{V}-10
s_yC^{V}-11	RPL8Ap-UTRa11-mTFP1-ADH1t	6.64*	p_yC^{V}-11
s_yC^{V}-12	RPL8Ap-UTRa12-mTFP1-ADH1t	6.76*	p_yC^{V}-12
s_yC^{V}-13	RPL8Ap-UTRa13-mTFP1-ADH1t	5.16*	p_yC^{V}-13
s_yC^{V}-14	RPL8Ap-UTRa14-mTFP1-ADH1t	5.18*	p_yC^{V}-14
s_yC^{V}-15	RPL8Ap-UTRa15-mTFP1-ADH1t	4.77*	p_yC^{V}-15
s_yC^{V}-16	RPL8Ap-UTRa16-mTFP1-ADH1t	4.89*	p_yC^{V}-16
s_yC^{VI}-1	TEF1p-UTRt1-*RcTal1*-ADH1t	2.71	p_yC^{VI}-1

s_yC^{VI}-2	TEF1p-UTRt2-*RcTal1*-ADH1t	3.43	p_yC^{VI}-2
s_yC^{VI}-3	TEF1p-UTRt3-*RcTal1*-ADH1t	4.70	p_yC^{VI}-3
s_yC^{VI}-4	TEF1p-UTRt4-*RcTal1*-ADH1t	6.89	p_yC^{VI}-4

Table S.2.2: Overview of the 5'UTR sequences generated by the PLS regression model and used in this study. UTR_RPL8A and UTR_TEF1 represent the native 5'UTRs of the *RPL8A* and *TEF1* gene respectively. The altered 10 bp parts of the 5'UTRs are presented in bold.

5'UTR name	5'UTR sequence
UTR_RPL8A	AAAACAACTAATTCGAA
UTRa1	AAAACAA**ACGCCTCAAA**
UTRa2	AAAACAA**ATGCCTCAAA**
UTRa3	AAAACAA**ACGCGTCAAA**
UTRa4	AAAACAA**ATGCGTCAAA**
UTRa5	AAAACAA**ACGCCTCACA**
UTRa6	AAAACAA**ATGCCTCACA**
UTRa7	AAAACAA**ACGCGTCACA**
UTRa8	AAAACAA**ATGCGTCACA**
UTRa9	AAAACAA**TCAACGAAAA**
UTRa10	AAAACAA**TCTACGAAAA**
UTRa11	AAAACAA**TCAAGGAAAA**
UTRa12	AAAACAA**TCTAGGAAAA**
UTRa13	AAAACAA**TCAACGATAA**
UTRa14	AAAACAA**TCTACGATAA**
UTRa15	AAAACAA**TCAAGGATAA**
UTRa16	AAAACAA**TCTAGGATAA**
UTRb1	AAAACAA**AGATCTAAAA**
UTRb2	AAAACAA**AGATGTAAAA**
UTRb3	AAAACAA**AGATCTACAA**
UTRb4	AAAACAA**AGATGTACAA**
UTRb5	AAAACAA**AGATCTAAAG**
UTRb6	AAAACAA**AGATGTAAAG**
UTRb7	AAAACAA**AGATCTACAG**
UTRb8	AAAACAA**AGATGTACAG**
UTRb9	AAAACAA**TAAGTGTAAA**
UTRb10	AAAACAA**TGAGTGTAAA**
UTRb11	AAAACAA**TAAGTGTACA**
UTRb12	AAAACAA**TGAGTGTACA**
UTRb13	AAAACAA**TAAGTGTAAC**
UTRb14	AAAACAA**TGAGTGTAAC**
UTRb15	AAAACAA**TAAGTGTACC**
UTRb16	AAAACAA**TGAGTGTACC**
UTR_TEF1	GCATAGCAATCTAATCTAAGTTTTAATTACAAA
UTRc1	GCATAGCAATCTAATCTAAGTTT**ACGGTATAAA**
UTRc2	GCATAGCAATCTAATCTAAGTTT**ACTGTATAAA**
UTRc3	GCATAGCAATCTAATCTAAGTTT**ACGGTATTAA**

UTRc4	GCATAGCAATCTAATCTAAGTTT**ACTGTATTAA**
UTRc5	GCATAGCAATCTAATCTAAGTTT**ACGGTATAAG**
UTRc6	GCATAGCAATCTAATCTAAGTTT**ACTGTATAAG**
UTRc7	GCATAGCAATCTAATCTAAGTTT**ACGGTATTAG**
UTRc8	GCATAGCAATCTAATCTAAGTTT**ACTGTATTAG**
UTRc9	GCATAGCAATCTAATCTAAGTTT**AGATCGTAAA**
UTRc10	GCATAGCAATCTAATCTAAGTTT**ATATCGTAAA**
UTRc11	GCATAGCAATCTAATCTAAGTTT**AGATCGTTAA**
UTRc12	GCATAGCAATCTAATCTAAGTTT**ATATCGTTAA**
UTRc13	GCATAGCAATCTAATCTAAGTTT**AGATCGTAAT**
UTRc14	GCATAGCAATCTAATCTAAGTTT**ATATCGTAAT**
UTRc15	GCATAGCAATCTAATCTAAGTTT**AGATCGTTAT**
UTRc16	GCATAGCAATCTAATCTAAGTTT**ATATCGTTAT**
UTRt1	GCATAGCAATCTAATCTAAGTTT**CGGATTCACCA**
UTRt2	GCATAGCAATCTAATCTAAGTTT**CGGATTCACAA**
UTRt3	GCATAGCAATCTAATCTAAGTTT**CGGATTCAAAA**
UTRt4	GCATAGCAATCTAATCTAAGTTT**AAAAAAAAAAA**

Table S.2.3: Definitions of all 13 features used in the Partial Least Square (PLS) regression model to predict protein abundances. All 13 features were obtained from the study of Dvir *et al.* [23] and were categorized in four main groups: AUG context, short k-mer sequences, uAUG's and RNA secondary structure (RSS). The adenine of the start codon (AUG) is position +1, all preceding nucleotides are numbered relative to this adenine ending with position -1 for the nucleotide in front of the start codon. A: adenine, T: thymine, G: guanine, C: cytosine.

Feature name	Definition	Category
AG_in_min3	The presence of an A or G at position -3.	AUG context
U_in_min3	The presence of a T at position -3.	
A_in_min1	The presence of an adenine at position -1.	
AA_in_min32	The presence of an AA motif at position [-3, -2].	
CG_in_min32	The presence of a CG motif at position [-3, -2].	
AC_in_min21	The presence of an AC motif at position [-2, -1].	
GACA_kmer	The presence of a GACA motif in the 5'UTR sequence.	Short k-mer sequences
GG_kmer	The presence of a GG motif in the 5'UTR sequence.	
CACC_kmer	The presence of a CACC motif in the 5'UTR sequence.	
CA_in_min76	The presence of a CA motif at position [-7, -6].	
CC_in_min76	The presence of a CC motif at position [-7, -6].	
oof_uAUG	The number of out-of-frame uAUG's in the 5'UTR sequence.	uAUG's
dG_EFE	The ensemble free energy (EFE). The EFE is calculated using RNAfold [243] and sums the Boltzmann weighted free energies of possible secondary structures of a given RNA sequence. To calculate the EFE, the whole 5'UTR and the first 50 nucleotides of the CDS were taken into account.	RSS

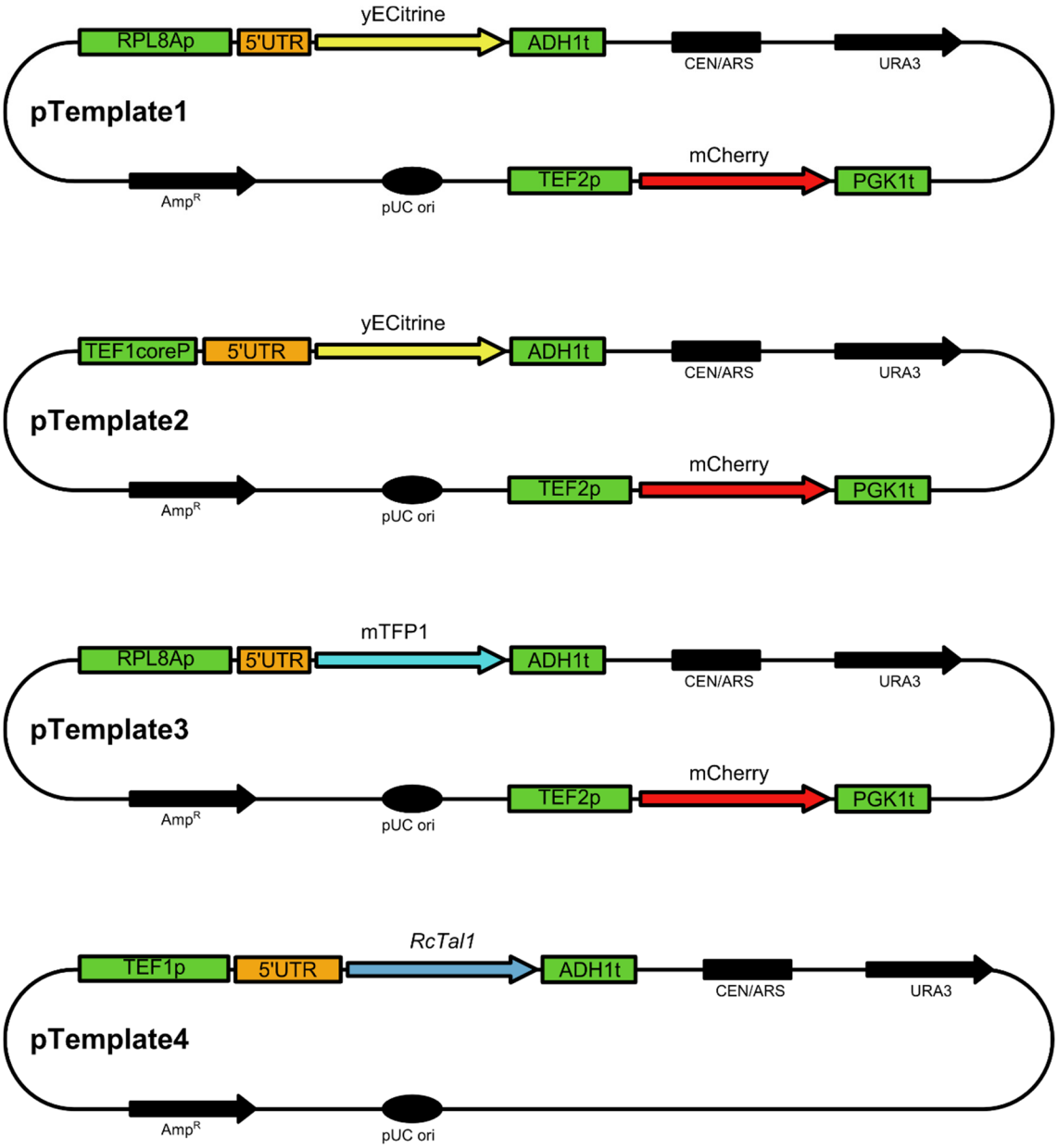

Figure S.2.1: Schematic overview of the pTemplate plasmids in this study. pTemplate plasmids were used for the amplification of linear DNA for the construction of the different p_yC expression vectors. pTemplate plasmids consist of a yeast low copy backbone containing a CEN/ARS ori and URA3 auxotrophic marker. In addition, an ampicillin resistance gene and pUC ori is present to maintain the plasmids in E. coli. pTemplate1 comprises the RPL8A promoter with its native 5'UTR in front of the yECitrine reporter and ADH1 terminator. pTemplate2 contains the TEF1 core promoter with its native 5'UTR in front of the yECitrine reporter and ADH1 terminator. pTemplate3 exists of the RPL8A promoter with its native 5'UTR in front of the mTFP1 reporter and ADH1 terminator. pTemplate4 contains the TEF1 promoter with its native 5'UTR in front of the *RcTal1* coding sequence and ADH1 terminator. All pTemplate plasmids, except pTemplate4, contain a mCherry transcription unit controlled by the TEF2 promoter and PGK1 terminator to correct for cellular background variation.

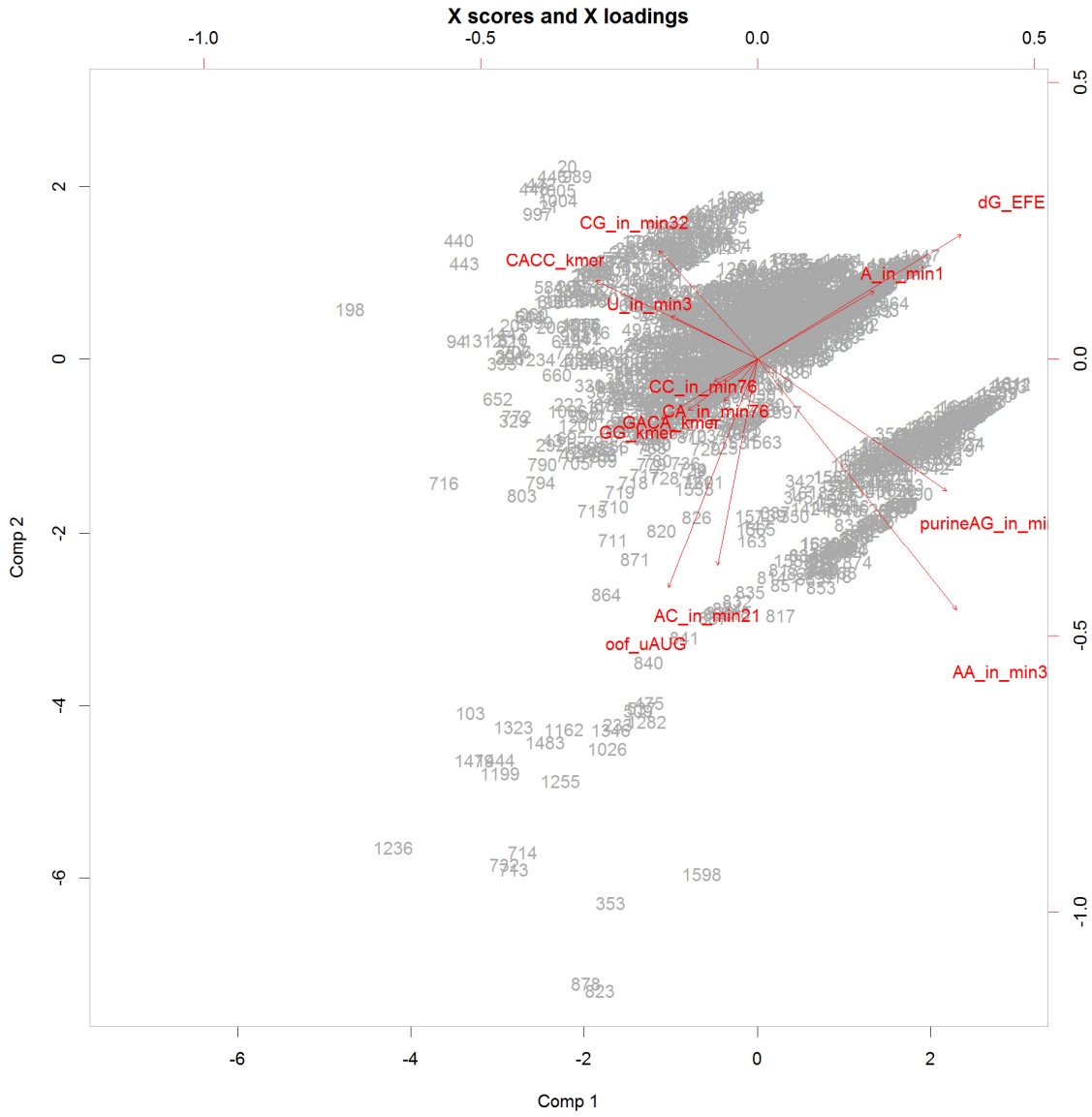

Figure S.2.2: Biplot of the first two components of the PLS regression model. An explanation of all features (AG_in_min3, U_in_min3, A_in_min1, AA_in_min32, CG_in_min32, AC_in_min21, GACA_kmer, GG_kmer, CACC_kmer, CA_in_min76, CC_in_min76, oof_uAUG, and dG_EFE) is given in Supplementary Table S.2.3.

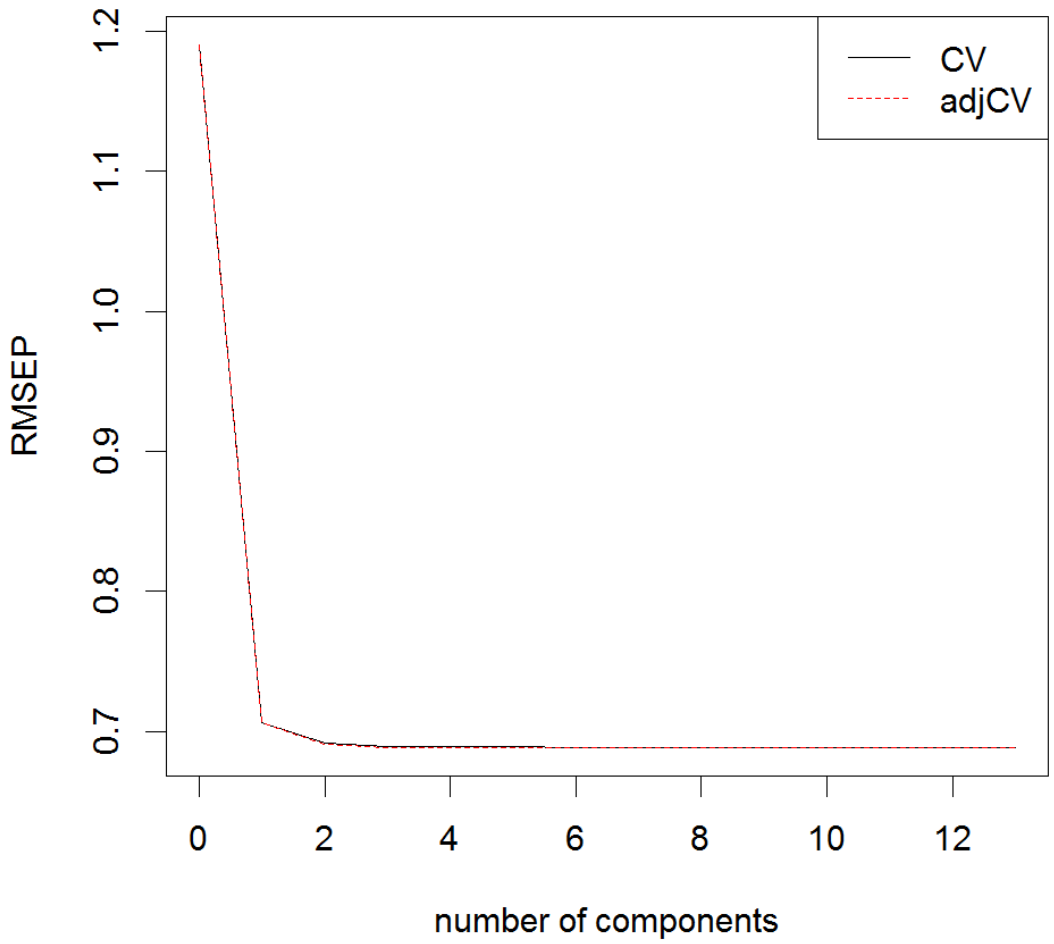

Figure S.2.3: Cross-validated root mean squared error of prediction (RMSEP) curves. CV is the ordinary cross-validation estimate, adjCV is a bias-corrected cross-validation estimate. From 4 components (*i.e.* latent variables), no further decrease in RMSEP was observed.

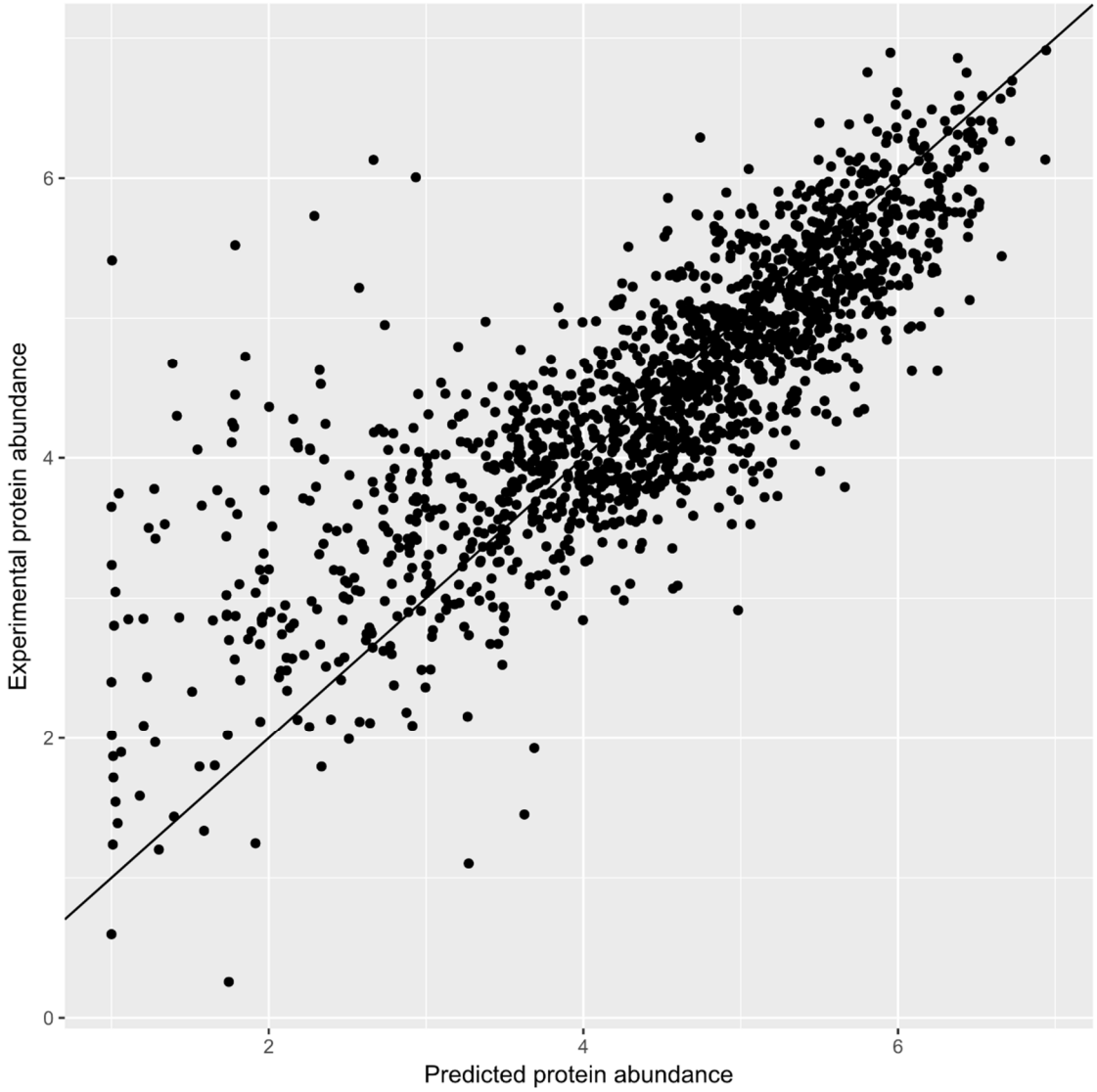

Figure S.2.4: Validation of the PLS regression model on the training set used for model calibration. The model uses 13 features of the 5'UTR of *Saccharomyces cerevisiae* (Supplementary Table S.2.3) to predict protein abundance. This plot represents the experimental [23] versus the predicted protein abundance, calculated via the PLS model, for the training set of 1633 5'UTRs. A coefficient of determination (R^2) of 0.67 was obtained.

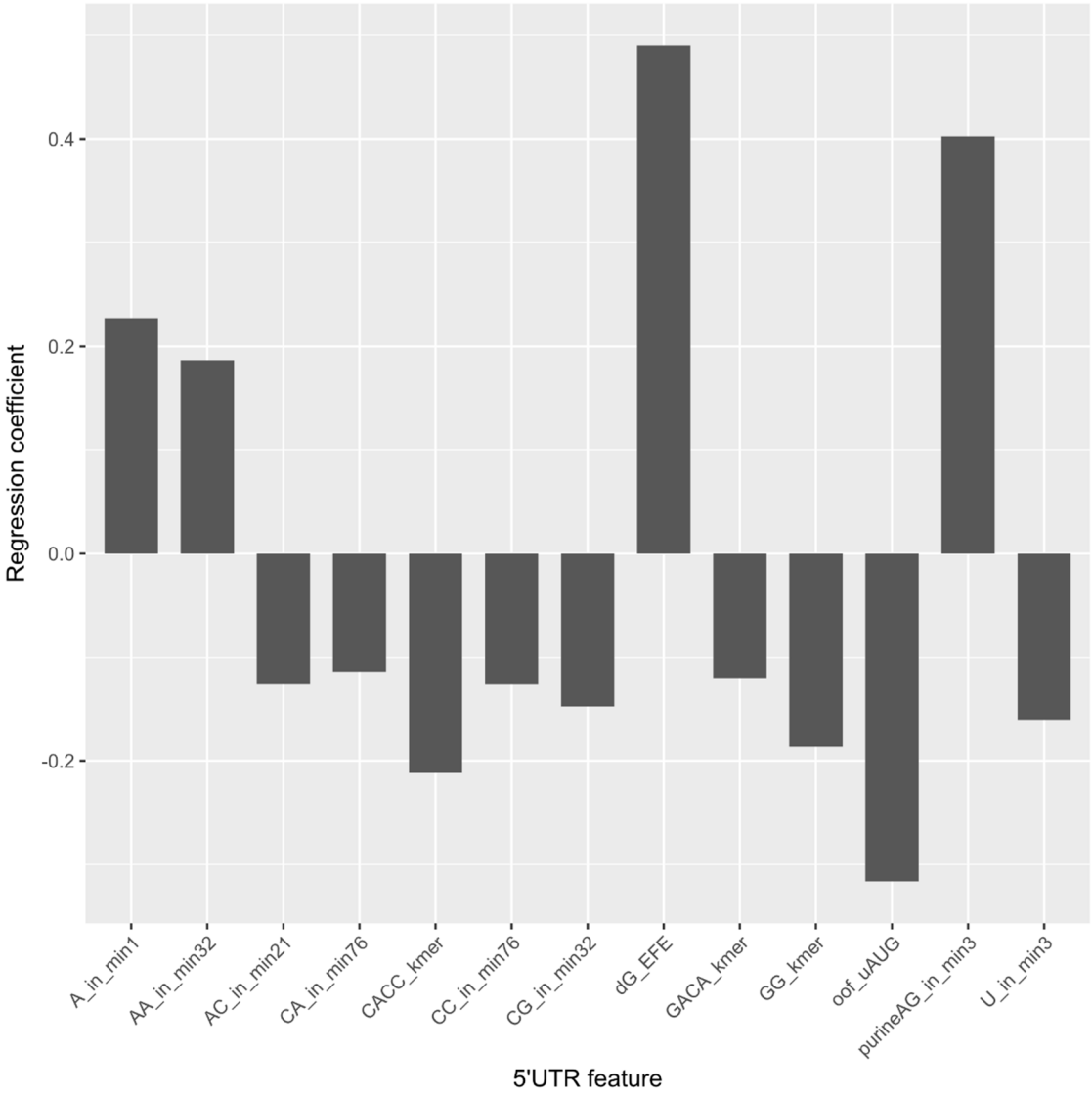

Figure S.2.5: The estimated regression coefficients of all 5'UTR features. An explanation of all features (AG_in_min3, U_in_min3, A_in_min1, AA_in_min32, CG_in_min32, AC_in_min21, GACA_kmer, GG_kmer, CACC_kmer, CA_in_min76, CC_in_min76, oof_uAUG, and dG_EFE) is available in Supplementary Table S.2.3.

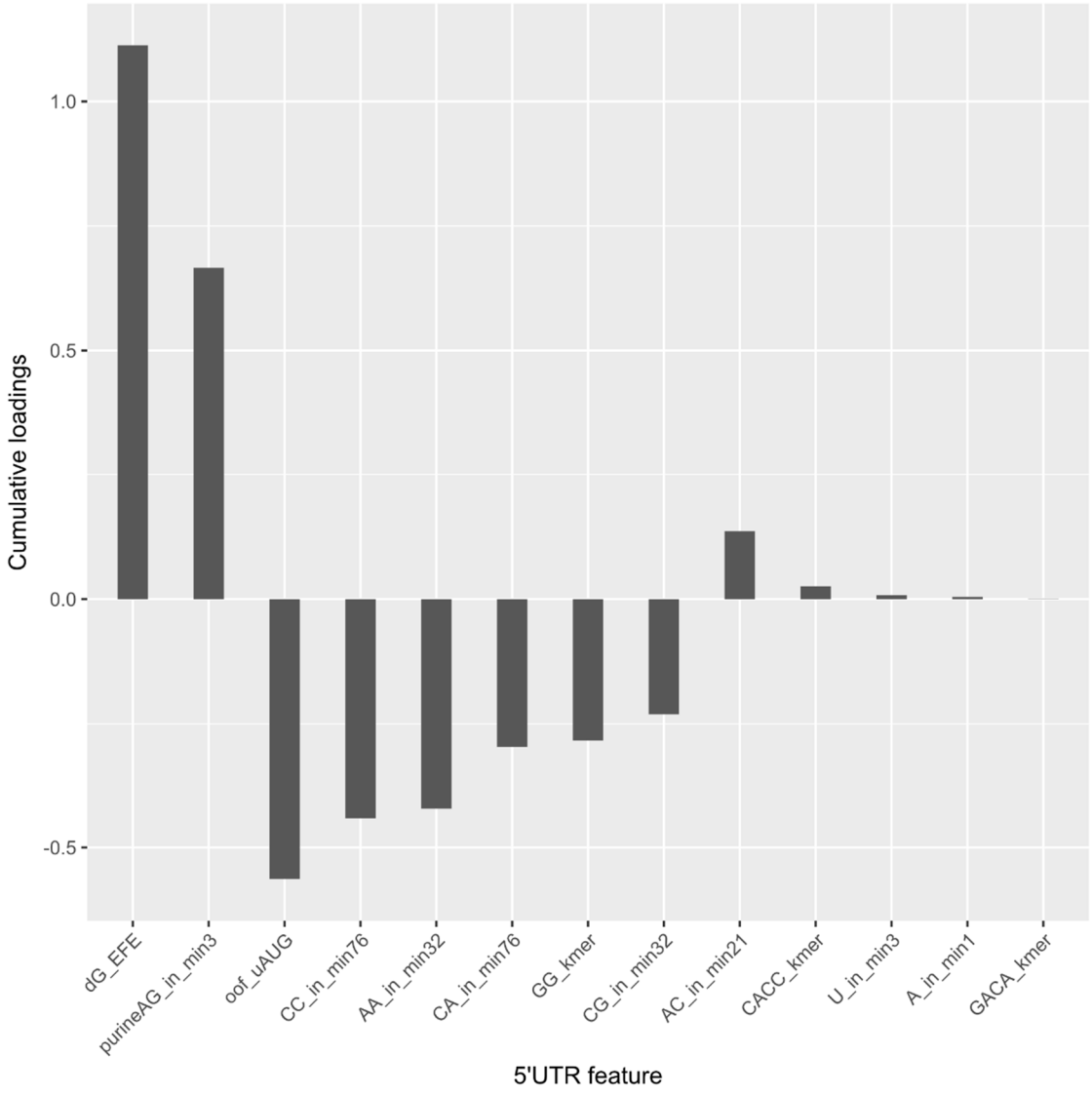

Figure S.2.6: Cumulative loadings of the four components used in the PLS model. An explanation of all features (AG_in_min3, U_in_min3, A_in_min1, AA_in_min32, CG_in_min32, AC_in_min21, GACA_kmer, GG_kmer, CACC_kmer, CA_in_min76, CC_in_min76, oof_uAUG, and dG_EFE) is available in Supplementary Table S.2.3.

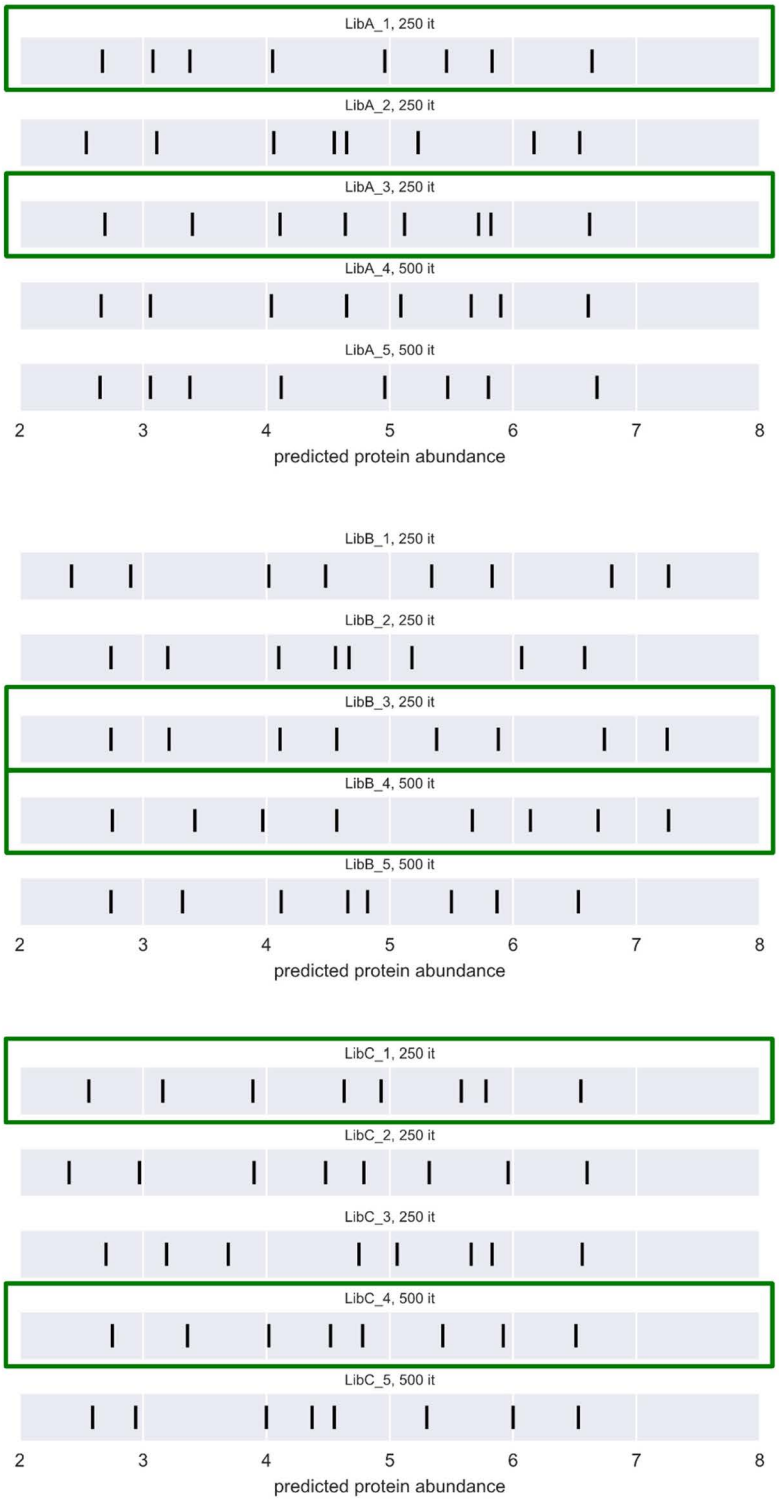

Figure S.2.7: Event plot representing the distribution of the predicted protein abundances for the calculated 5'UTR libraries (UTRa, UTRb and UTRc). Respectively libA_1 & libA_3, libB_3 & libB_4 and libC_1 & libC_4 were selected (indicated by a green box) and form library UTRa, UTRb and UTRc. All three libraries are available in Supplementary Table S.2.2.

```
LOCUS       linearDNA-p_yUTR        1278 bp    DNA     linear   SYN 05-APR-2017
DEFINITION  p_yUTRA cut 7094 to 957
ACCESSION   linearDNA-p_yUTR
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown
            Unclassified.
REFERENCE   1  (bases 1 to 1278)
  AUTHORS   Self
  JOURNAL   Unpublished.
COMMENT     SECID/File created by Clone Manager, Scientific & Educational Software
FEATURES             Location/Qualifiers
     misc_feature    1..334
                     /gene="pRPL8A"
                     /product="RPL8A promoter"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    335..351
                     /gene="RPL8A 5'UTR"
                     /product="17 bp 5'UTR of the RPL8A gene"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_signal     342..351
                     /label=UTRa
                     /SECDrawAs="Label"
     CDS             352..1068
                     /gene="'yECitrine"
                     /product="yeast enhanced yellow fluorescent protein"
                     /codon_start=2
                     /translation="CLKVKNYSLVLSQFWLN"
                     /SECDrawAs="Gene"
                     /SECStyleId=1
                     /SECName="yECitrine"
                     /SECDescr="yeast enhanced yellow fluorescent protein"
     misc_feature    1076..1278
                     /gene="ADH1t"
                     /product="ADH1 terminator"
                     /SECDrawAs="Region"
                     /SECStyleId=1
ORIGIN
        1 acataaataa tttctattaa caatgtaatt tccataattt tatattcctc tccaccttct
       61 attgcatcat gtactattca aatgactgta acactagtat tatgaagaaa acacccaaac
      121 atatctaggc catcagattt ttttttttc atttttcatt ttttctctcat tttcttattt
      181 atttttattg aaaaataata accgacgcaa acaaattgga aaaaccaacg caaaaaaaaa
      241 aagacgctaa attgtttata aaggcgagga atttgtatct atcaattact attccagttg
      301 tcagtttaca ttgcttaccc tctattatca catcaaaaca annnnnnnnn natgtctaaa
      361 ggtgaagaat tattcactgg tgttgtccca attttggttg aattagatgg tgatgttaat
      421 ggtcacaaat tttctgtctc cggtgaaggt gaaggtgatg ctacttacgg taaattgacc
      481 ttaaaattta tttgtactac tggtaaattg ccagttccat ggccaacctt agtcactact
      541 ttaggttatg gtttgatgtg ttttgctaga tacccagatc atatgaaaca acatgacttt
      601 ttcaagtctg ccatgccaga aggttatgtt caagaaagaa ctatttttttt caaagatgac
      661 ggtaactaca agaccagagc tgaagtcaag tttgaaggtg ataccttagt taatagaatc
      721 gaattaaaag gtattgattt taaagaagat ggtaacattt taggtcacaa attggaatac
      781 aactataact ctcacaatgt ttacatcatg gctgacaaac aaaagaatgg tatcaaagtt
      841 aacttcaaaa ttagacacaa cattgaagat ggttctgttc aattagctga ccattatcaa
      901 caaaatactc caattggtga tggtccagtc ttgttaccag acaaccatta cttatcctat
      961 caatctgcct tatccaaaga tccaaacgaa aagagagacc acatggtctt gttagaattt
     1021 gttactgctg ctggtattac ccatggtatg gatgaattgt acaaataagg cgcgccactt
     1081 ctaaataagc gaatttctta tgatttatga tttttattat taaataagtt ataaaaaaaa
     1141 taagtgtata caaattttaa agtgactctt aggttttaaa acgaaaattc ttattcttga
     1201 gtaactcttt cctgtaggtc aggttgcttt ctcaggtata gtatgaggtc gctcttattg
     1261 accacacctc taccggca
//
```

Figure S.2.8: Annotated Genbank file of the *RPL8Ap-UTRai-yECitrine-ADH1t* transcription unit in expression vector p_yC^{I}-i (i varies from 1 to 16). The 5'UTR is underlined and indicated in bold. The respective sequences of library UTRa are represented in Supplementary Table S.2.2.

```
LOCUS       linearDNA-p_yUTR        1120 bp    DNA     linear   SYN 05-APR-2017
DEFINITION  p2a_TEF1coreP-RPL8A_UTRlib1_G1-yECit cut 1 to 1120
ACCESSION   linearDNA-p_yUTR
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown
            Unclassified.
REFERENCE   1  (bases 1 to 1120)
  AUTHORS   Self
  JOURNAL   Unpublished.
COMMENT     SECID/File created by Clone Manager, Scientific & Educational Software
FEATURES             Location/Qualifiers
     misc_feature    1..176
                     /gene="cpTEF1"
                     /product="TEF1 core promoter"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    177..193
                     /gene="RPL8A 5'UTR"
                     /product="17 bp 5'UTR of the RPL8A gene"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_signal     184..193
                     /label=UTRa
                     /SECDrawAs="Label"
     CDS             194..910
                     /gene="'yECitrine"
                     /product="yeast enhanced yellow fluorescent protein"
                     /codon_start=3
                     /SECDrawAs="Gene"
                     /SECStyleId=1
                     /SECName="yECitrine"
                     /SECDescr="yeast enhanced yellow fluorescent protein"
     misc_feature    918..1120
                     /gene="ADH1t"
                     /product="ADH1 terminator"
                     /SECDrawAs="Region"
                     /SECStyleId=1
ORIGIN
        1 aataaaaatt tttatcacgt ttctttttct tgaaaatttt tttttttgat ttttttctct
       61 ttcgatgacc tcccattgat atttaagtta ataaacggtc ttcaatttct caagtttcag
      121 tttcattttt cttgttctat tacaactttt tttacttctt gctcattaga aagaaaaaaa
      181 caannnnnnn nnnatgtcta aaggtgaaga attattcact ggtgttgtcc caattttggt
      241 tgaattagat ggtgatgtta atggtcacaa attttctgtc tccggtgaag gtgaaggtga
      301 tgctacttac ggtaaattga ccttaaaatt tatttgtact actggtaaat tgccagttcc
      361 atggccaacc ttagtcacta ctttaggtta tggtttgatg tgttttgcta gatacccaga
      421 tcatatgaaa caacatgact ttttcaagtc tgccatgcca gaaggttatg ttcaagaaag
      481 aactattttt ttcaaagatg acggtaacta caagaccaga gctgaagtca agtttgaagg
      541 tgatacctta gttaatagaa tcgaattaaa aggtattgat tttaaagaag atggtaacat
      601 tttaggtcac aaattggaat acaactataa ctctcacaat gtttacatca tggctgacaa
      661 acaaaagaat ggtatcaaag ttaacttcaa aattagacac aacattgaag atggttctgt
      721 tcaattagct gaccattatc aacaaaatac tccaattggt gatggtccag tcttgttacc
      781 agacaaccat tacttatcct atcaatctgc cttatccaaa gatccaaacg aaaagagaga
      841 ccacatggtc ttgttagaat ttgttactgc tgctggtatt acccatggta tggatgaatt
      901 gtacaaataa ggcgcgccac ttctaaataa gcgaatttct tatgatttat gatttttatt
      961 attaaataag ttataaaaaa aataagtgta tacaaatttt aaagtgactc ttaggtttta
     1021 aaacgaaaat tcttattctt gagtaactct ttcctgtagg tcaggttgct ttctcaggta
     1081 tagtatgagg tcgctcttat tgaccacacc tctaccggca
//
```

Figure S.2.9: Annotated Genbank file of the *TEF1coreP-UTRai-yECitrine-ADH1t* transcription unit in expression vector p_yC^{II}-i (i varies from 1 to 16). The 5'UTR is underlined and indicated in bold. The respective sequences of library UTRa are represented in Supplementary Table S.2.2.

```
LOCUS       linearDNA-p_yUTR        1272 bp    DNA     linear   SYN 05-APR-2017
DEFINITION  p2a_RPL8Ap-RPL8A-libB-mTFP1_1 cut 7068 to 931
ACCESSION   linearDNA-p_yUTR
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown
            Unclassified.
REFERENCE   1  (bases 1 to 1272)
  AUTHORS   Self
  JOURNAL   Unpublished.
COMMENT     SECID/File created by Clone Manager, Scientific & Educational Software
FEATURES             Location/Qualifiers
     misc_feature    1..334
                     /gene="pRPL8A"
                     /product="RPL8A promoter"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    335..351
                     /gene="RPL8A 5'UTR"
                     /product="17 bp 5'UTR of the RPL8A gene"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_signal     342..351
                     /label=UTRb
                     /SECDrawAs="Label"
     CDS             352..1062
                     /gene="mTFP1_co_Sc"
                     /product="yeast codon optimized mTFP1 fluorescent protein"
                     /SECDrawAs="Gene"
                     /SECStyleId=1
                     /SECName="mTFP1_co_Sc"
                     /SECDescr="yeast codon optimized mTFP1 fluorescent protein"
     misc_feature    1070..1272
                     /gene="ADH1t"
                     /product="ADH1 terminator"
                     /SECDrawAs="Region"
                     /SECStyleId=1
ORIGIN
        1 acataaataa tttctattaa caatgtaatt tccataattt tatattcctc tccaccttct
       61 attgcatcat gtactattca aatgactgta acactagtat tatgaagaaa acacccaaac
      121 atatctaggc catcagattt ttttttttc atttttcatt tttttctcat tttcttattt
      181 atttttattg aaaaataata accgacgcaa acaaattgga aaaaccaacg caaaaaaaaa
      241 aagacgctaa attgtttata aaggcgagga atttgtatct atcaattact attccagttg
      301 tcagtttaca ttgcttaccc tctattatca catcaaaaca annnnnnnnn natggtgagt
      361 aaggggaag aaactactat gggagtaatc aagcccgaca tgaagattaa gttaaagatg
      421 gaagggaacg tgaacggtca cgcattcgtt atcgagggag aaggagaagg caagccctat
      481 gatggcacaa atacgataaa tctggaagtg aaagaaggag cgcctctgcc tttttcctac
      541 gatatactga caacggcgtt tgcctacgga aacagggcgt tcaccaagta ccctgacgat
      601 atcccgaatt acttcaagca atcattccct gaaggatata gttgggagcg tacgatgacg
      661 tttgaggata agggaatagt caaggttaag tcagatatat ctatggaaga agattccttt
      721 atatatgaga tacatttgaa aggtgagaac ttccccccga atggtcctgt aatgcaaaaa
      781 aagaccactg ggtgggacgc gtctaccgag cgtatgtacg tcagagatgg ggtactaaaa
      841 ggagatgtga aacataagtt attattggag ggcggcggcc atcaccgtgt ggacttcaaa
      901 actatttata gagcgaaaaa agccgtgaag ctaccagatt atcattttgt agaccacaga
      961 atcgagattc tgaaccatga taaagactat aataaggtta ctgtgtatga gagcgccgtt
     1021 gcgaggaact ctactgacgg aatggatgaa ttatataaat aaggcgcgcc acttctaaat
     1081 aagcgaattt cttatgattt atgattttta ttattaaata agttataaaa aaaataagtg
     1141 tatacaaatt ttaaagtgac tcttaggttt taaaacgaaa attcttattc ttgagtaact
     1201 ctttcctgta ggtcaggttg ctttctcagg tatagtatga ggtcgctctt attgaccaca
     1261 cctctaccgg ca
//
```

Figure S.2.10: Annotated Genbank file of the *RPL8Ap-UTRbi-mTFP1-ADH1t* transcription unit in expression vector p_yC^{III}-i (i varies from 1 to 16). The 5'UTR is underlined and indicated in bold. The respective sequences of library UTRb are represented in Supplementary Table S.2.2.

```
LOCUS       linearDNA-p_yUTR        1136 bp    DNA     linear   SYN 05-APR-2017
DEFINITION  p2a_cTEF1p_UTR-TEF1-libC-yECit_1 cut 7074 to 937
ACCESSION   linearDNA-p_yUTR
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown
            Unclassified.
REFERENCE   1  (bases 1 to 1136)
  AUTHORS   Self
  JOURNAL   Unpublished.
COMMENT     SECID/File created by Clone Manager, Scientific & Educational Software
FEATURES             Location/Qualifiers
     misc_feature    1..176
                     /gene="cpTEF1"
                     /product="TEF1 core promoter"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    177..209
                     /gene="TEF1 5'UTR"
                     /product="33 bp 5'UTR of the TEF1 gene"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_signal     200..209
                     /label=UTRc
                     /SECDrawAs="Label"
     CDS             210..926
                     /gene="'yECitrine"
                     /product="yeast enhanced yellow fluorescent protein"
                     /codon_start=2
                     /SECDrawAs="Gene"
                     /SECStyleId=1
                     /SECName="yECitrine"
                     /SECDescr="yeast enhanced yellow fluorescent protein"
     misc_feature    934..1136
                     /gene="ADH1t"
                     /product="ADH1 terminator"
                     /SECDrawAs="Region"
                     /SECStyleId=1
ORIGIN
        1 aataaaaatt tttatcacgt ttctttttct tgaaaatttt tttttttgat ttttttctct
       61 ttcgatgacc tcccattgat atttaagtta ataaacggtc ttcaatttct caagtttcag
      121 tttcattttt cttgttctat tacaactttt tttacttctt gctcattaga aagaaa gcat
      181 agcaatctaa tctaagtttn nnnnnnnnna tgtctaaagg tgaagaatta ttcactggtg
      241 ttgtcccaat tttggttgaa ttagatggtg atgttaatgg tcacaaattt tctgtctccg
      301 gtgaaggtga aggtgatgct acttacggta aattgacctt aaaatttatt tgtactactg
      361 gtaaattgcc agttccatgg ccaaccttag tcactacttt aggttatggt ttgatgtgtt
      421 ttgctagata cccagatcat atgaaacaac atgacttttt caagtctgcc atgccagaag
      481 gttatgttca agaaagaact attttttca aagatgacgg taactacaag accagagctg
      541 aagtcaagtt tgaaggtgat accttagtta atagaatcga attaaaaggt attgatttta
      601 aagaagatgg taacatttta ggtcacaaat tggaatacaa ctataactct cacaatgttt
      661 acatcatggc tgacaaacaa aagaatggta tcaaagttaa cttcaaaatt agacacaaca
      721 ttgaagatgg ttctgttcaa ttagctgacc attatcaaca aaatactcca attggtgatg
      781 gtccagtctt gttaccagac aaccattact tatcctatca atctgcctta tccaaagatc
      841 caaacgaaaa gagagaccac atggtcttgt tagaatttgt tactgctgct ggtattaccc
      901 atggtatgga tgaattgtac aaataaggcg cgccacttct aaataagcga atttcttatg
      961 atttatgatt tttattatta aataagttat aaaaaaaata agtgtataca aattttaaag
     1021 tgactcttag gttttaaaac gaaaattctt attcttgagt aactctttcc tgtaggtcag
     1081 gttgctttct caggtatagt atgaggtcgc tcttattgac cacacctcta ccggca
//
```

Figure S.2.11: Annotated Genbank file of the *TEF1coreP-UTRci-yECitrine-ADH1t* transcription unit in expression vector p_yC^{IV}-i (i varies from 1 to 16). The 5'UTR is underlined and indicated in bold. The respective sequences of library UTRc are represented in Supplementary Table S.2.2.

```
LOCUS       linearDNA-p_yUTR        1272 bp    DNA     linear   SYN 05-APR-2017
DEFINITION  p2a_RPL8Ap-RPL8AUTR-libA_G1 cut 4125 to 5397
ACCESSION   linearDNA-p_yUTR
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown
            Unclassified.
REFERENCE   1  (bases 1 to 1272)
  AUTHORS   Self
  JOURNAL   Unpublished.
COMMENT     SECID/File created by Clone Manager, Scientific & Educational Software
FEATURES             Location/Qualifiers
     misc_feature    1..341
                     /gene="pRPL8A"
                     /product="RPL8A promoter"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    335..351
                     /gene="RPL8A 5'UTR"
                     /product="17 bp 5'UTR of the RPL8A gene"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_signal     342..351
                     /label=UTRa
                     /SECDrawAs="Label"
     CDS             352..1062
                     /gene="mTFP1_co_Sc"
                     /product="yeast codon optimized mTFP1 fluorescent protein"
                     /SECDrawAs="Gene"
                     /SECStyleId=1
                     /SECName="mTFP1_co_Sc"
                     /SECDescr="yeast codon optimized mTFP1 fluorescent protein"
     misc_feature    1070..1272
                     /gene="ADH1t"
                     /product="ADH1 terminator"
                     /SECDrawAs="Region"
                     /SECStyleId=1
ORIGIN
        1 acataaataa tttctattaa caatgtaatt tccataattt tatattcctc tccaccttct
       61 attgcatcat gtactattca aatgactgta acactagtat tatgaagaaa acacccaaac
      121 atatctaggc catcagattt ttttttttc atttttcatt tttttctcat tttcttattt
      181 atttttattg aaaaataata accgacgcaa acaaattgga aaaaccaacg caaaaaaaaa
      241 aagacgctaa attgtttata aaggcgagga atttgtatct atcaattact attccagttg
      301 tcagtttaca ttgcttaccc tctattatca catcaaaaca annnnnnnnn natggtgagt
      361 aaggggaag aaactactat gggagtaatc aagcccgaca tgaagattaa gttaaagatg
      421 gaagggaacg tgaacggtca cgcattcgtt atcgagggag aaggagaagg caagccctat
      481 gatggcacaa atacgataaa tctggaagtg aaagaaggag cgcctctgcc tttttcctac
      541 gatatactga caacggcgtt tgcctacgga aacagggcgt tcaccaagta ccctgacgat
      601 atcccgaatt acttcaagca atcattccct gaaggatata gttgggagcg tacgatgacg
      661 tttgaggata agggaatagt caaggttaag tcagatatat ctatggaaga agattccttt
      721 atatatgaga tacatttgaa aggtgagaac ttccccccga atggtcctgt aatgcaaaaa
      781 aagaccactg ggtgggacgc gtctaccgag cgtatgtacg tcagagatgg ggtactaaaa
      841 ggagatgtga aacataagtt attattggag ggcggcggcc atcaccgtgt ggacttcaaa
      901 actatttata gagcgaaaaa agccgtgaag ctaccagatt atcattttgt agaccacaga
      961 atcgagattc tgaaccatga taaagactat aataaggtta ctgtgtatga gagcgccgtt
     1021 gcgaggaact ctactgacgg aatggatgaa ttatataaat aaggcgcgcc acttctaaat
     1081 aagcgaattt cttatgattt atgattttta ttattaaata agttataaaa aaaataagtg
     1141 tatacaaatt ttaaagtgac tcttaggttt taaaacgaaa attcttattc ttgagtaact
     1201 ctttcctgta ggtcaggttg ctttctcagg tatagtatga ggtcgctctt attgaccaca
     1261 cctctaccgg ca
//
```

Figure S.2.12: Annotated Genbank file of the *RPL8Ap-UTRai-mTFP1-ADH1t* transcription unit in expression vector p_yC^{V}-i (i varies from 1 to 16). The 5'UTR is underlined and indicated in bold. The respective sequences of library UTRa are represented in Supplementary Table S.2.2.

```
LOCUS       TEF1p-RcTal1-ADH        2219 bp    DNA     linear   SYN 13-DEC-2017
DEFINITION  p2a33_5UTRnative_TAL1 cut 2590 to 4808
ACCESSION   TEF1p-RcTal1-ADH
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown
            Unclassified.
REFERENCE   1  (bases 1 to 2219)
  AUTHORS   Self
  JOURNAL   Unpublished.
COMMENT     SECID/File created by Clone Manager, Scientific & Educational Software
FEATURES             Location/Qualifiers
     misc_feature    1..203
                     /gene="UAS_TEF1"
                     /product="upstream activating sequence TEF1 gene"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    204..413
                     /gene="TEF1 core promoter"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_feature    380..413
                     /gene="TEF1 5'UTR"
                     /SECDrawAs="Region"
                     /SECStyleId=1
     misc_signal     403..413
                     /label=RcTal1-UTRlib
                     /SECDrawAs="Label"
     CDS             414..2009
                     /gene="co_Rc_TAL1'"
                     /SECDrawAs="Gene"
                     /SECStyleId=1
                     /SECName="Rc_coSc_TAL1"
     misc_feature    2017..2219
                     /gene="ADH1t"
                     /product="ADH terminator"
                     /SECDrawAs="Region"
                     /SECStyleId=1
ORIGIN
        1 atagcttcaa aatgtttcta ctccttttt actcttccag attttctcgg actccgcgca
       61 tcgccgtacc acttcaaaac acccaagcac agcatactaa atttcccctc tttcttcctc
      121 tagggtgtcg ttaattaccc gtactaaagg tttggaaaag aaaaaagaga ccgcctcgtt
      181 tctttttctt cgtcgaaaaa ggcaataaaa atttttatca cgtttctttt tcttgaaaat
      241 ttttttttt gatttttttc tctttcgatg acctcccatt gatatttaag ttaataaacg
      301 gtcttcaatt tctcaagttt cagtttcatt tttcttgttc tattacaact ttttttactt
      361 cttgctcatt agaaagaaag g catagcaatc taatctaagt ttnnnnnnnn nnnatgacct
      421 tacaatccca aactgccaaa gactgcttag ccttagacgg tgccttgacc ttggttcaat
      481 gtgaagcaat tgccacacat agatccagaa taagtgtcac cccagctttg agagaaagat
      541 gcgctagagc acatgccaga ttagaacacg ctattgcaga acaaagacac atctatggta
      601 taactacagg ttttggtcct ttggctaata gattaatagg tgccgatcaa ggtgctgaat
      661 tgcaacaaaa cttaatctac catttggcta ctggtgttgg tccaaaattg tcttgggccg
      721 aagctagagc attgatgttg gcaagattga actcaatctt gcaaggtgca tctggtgcct
      781 cacctgaaac aatcgacaga attgttgctg tcttaaacgc tggtttcgca ccagaagtcc
      841 ctgcccaagg tactgtaggt gcttccggtg acttgacacc attggcacat atggttttgg
      901 ccttacaagg tagaggtaga atgattgatc ctagtggtag agttcaagaa gccggtgctg
      961 tcatggacag attatgtggt ggtccattga ctttagctgc aagagatggt ttggctttag
     1021 ttaatggtac ttctgccatg acagctatcg ccgctttgac aggtgttgaa gcagccagag
     1081 ctattgatgc tgcattaaga cattccgcag tattaatgga agttttgagt ggtcatgcag
     1141 aagcctggca cccagctttt gcagaattaa gaccacaccc tggtcaatta agagctaccg
     1201 aaagattagc ccaagctttg gatggtgcag gtagagtttg cagaaccttg actgccgcta
     1261 gaagattgac agcagccgac ttaagaccag aagatcatcc tgcacaagac gcctattctt
     1321 tgagagttgt cccacaatta gttggtgctg tctgggatac tttggactgg cacgatagag
     1381 tagttacctg tgaattgaac tcagtcactg ataacccaat atttcctgaa ggttgcgctg
     1441 tacctgcatt acatggtggt aatttcatgg gtgtacacgt tgcattggcc tccgacgctt
     1501 taaacgctgc attagtaaca ttggctggtt tagttgaaag acaaatcgca agattgaccg
     1561 atgaaaagtt gaataagggt ttgccagcat ttttgcatgg tggtcaagca ggtttacaat
     1621 caggtttcat gggtgctcaa gttacagcta ccgcattgtt agcagaaatg agagccaacg
     1681 ctacccctgt ctctgtacaa tctttgtcaa ctaatggtgc taaccaagat gtcgtatcaa
     1741 tgggtactat cgccgctaga agagcaagag cccaattgtt gccattgtct caaatccaag
     1801 caatcttggc tttagcattg gcccaagcta tggacttgtt agatgaccct gaaggtcaag
```

```
     1861 caggttggtc cttgacagcc agagacttaa gagatagaat tagagctgtt agtccaggtt
     1921 tgagagctga tagaccttta gcaggtcata tagaagcagt cgcacaaggt ttgagacatc
     1981 catccgccgc agcagaccct ccagcctaag gcgcgccact tctaaataag cgaatttctt
     2041 atgatttatg atttttatta ttaaataagt tataaaaaaa ataagtgtat acaaatttta
     2101 aagtgactct taggttttaa aacgaaaatt cttattcttg agtaactctt tcctgtaggt
     2161 caggttgctt tctcaggtat agtatgaggt cgctcttatt gaccacacct ctaccggca
//
```

Figure S.2.13: Annotated Genbank file of the *TEF1p-UTRti-RcTal1-ADH1t* transcription unit in expression vector p_yC^{VI}-i (i varies from 1 to 4). The 5'UTR is underlined and indicated in bold. The respective sequences of library UTRt are represented in Supplementary Table S.2.2.

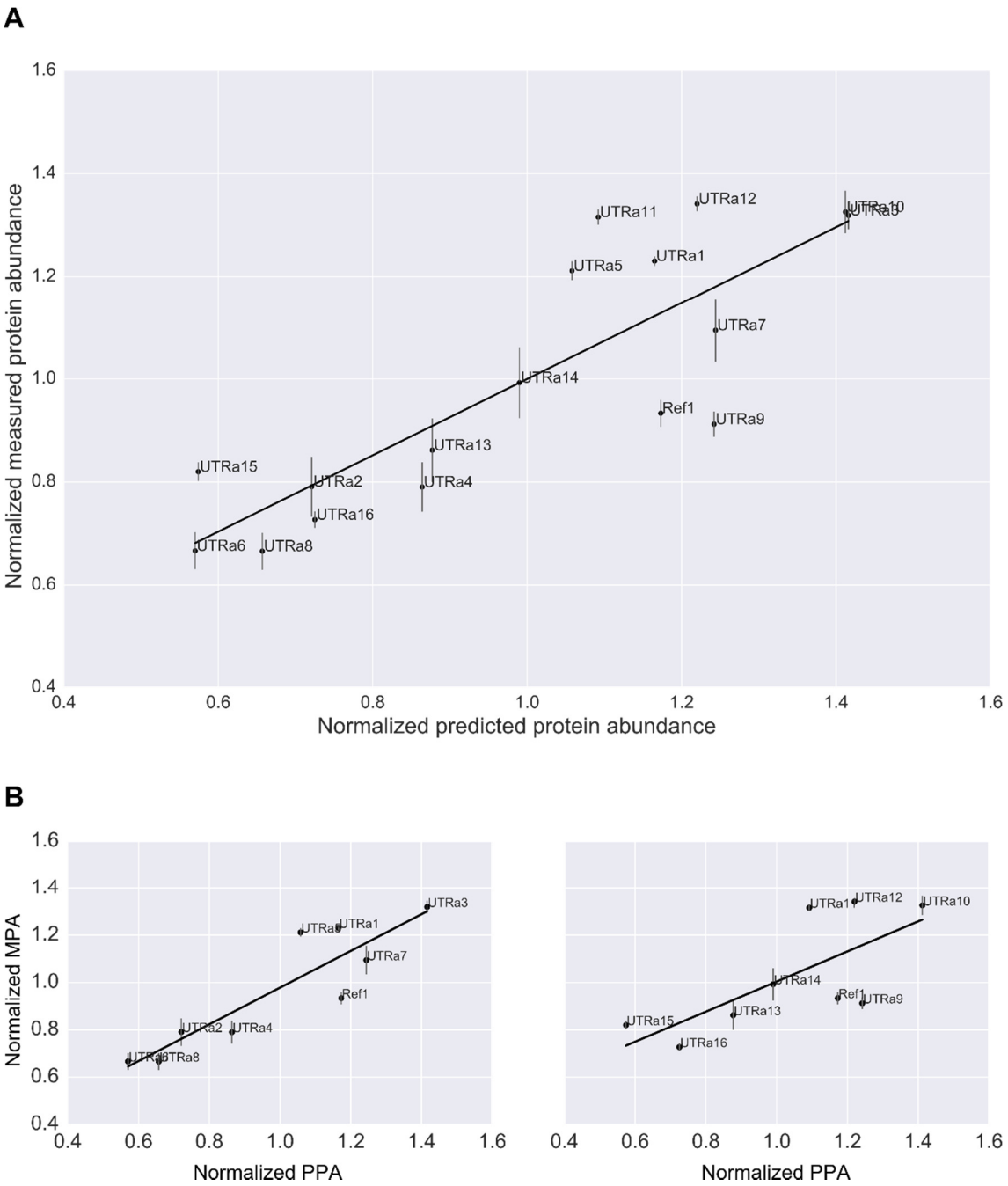

Figure S.2.14: OLS regression plots comparing the normalized predicted protein abundance (PPA) calculated by forward engineering with our model and the normalized measured protein abundance (MPA), determined by measuring yECitrine-to-mCherry ratios. (A) Regression plot of both calculated 8-containing 5'UTR libraries representing strains s_yC^{I}-1 to s_yC^{I}-16, additionally, reference strain sTemplate1 was included (R^2 = 0.70). (B) Left: Regression plot of the first part of library UTRa consisting of eight 5'UTR candidates representing strains s_yC^{I}-1 to s_yC^{I}-8 including reference strain sTemplate1 (R^2 = 0.81). Right: Regression plot of the second part of library UTRa consisting of eight 5'UTR candidates representing strains s_yC^{I}-9 to s_yC^{I}-16 including reference strain sTemplate1 (R^2 = 0.51). Error bars represent standard error of the mean of four biological replicates.

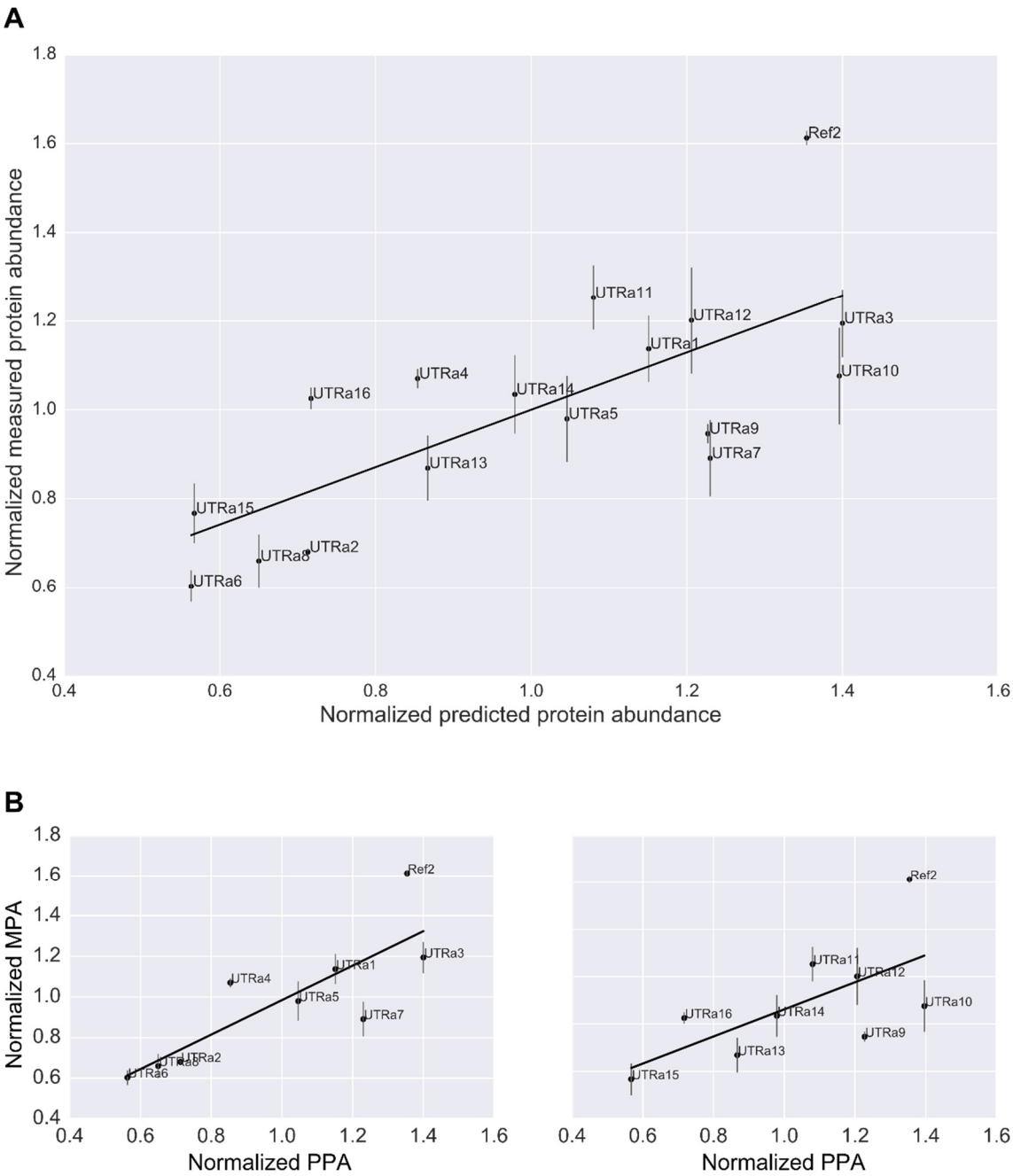

Figure S.2.15: OLS regression plots comparing the normalized predicted protein abundance (PPA) calculated by forward engineering with our model and the normalized measured protein abundance (MPA), determined by measuring yECitrine-to-mCherry ratios. (A) Regression plot of both calculated 8-containing 5'UTR libraries representing strains s_yCII-1 to s_yCII-16, additionally, reference strain sTemplate2 was included (R^2 = 0.54). (B) Left: Regression plot of the first part of library UTRa consisting of eight 5'UTR candidates representing strains s_yCII-1 to s_yCII-8 including reference strain sTemplate2 (R^2 = 0.69). Right: Regression plot of the second part of library UTRa consisting of eight 5'UTR candidates representing strains s_yCII-9 to s_yCII-16 including reference strain sTemplate2 (R^2 = 0.43). Error bars represent standard error of the mean of four biological replicates.

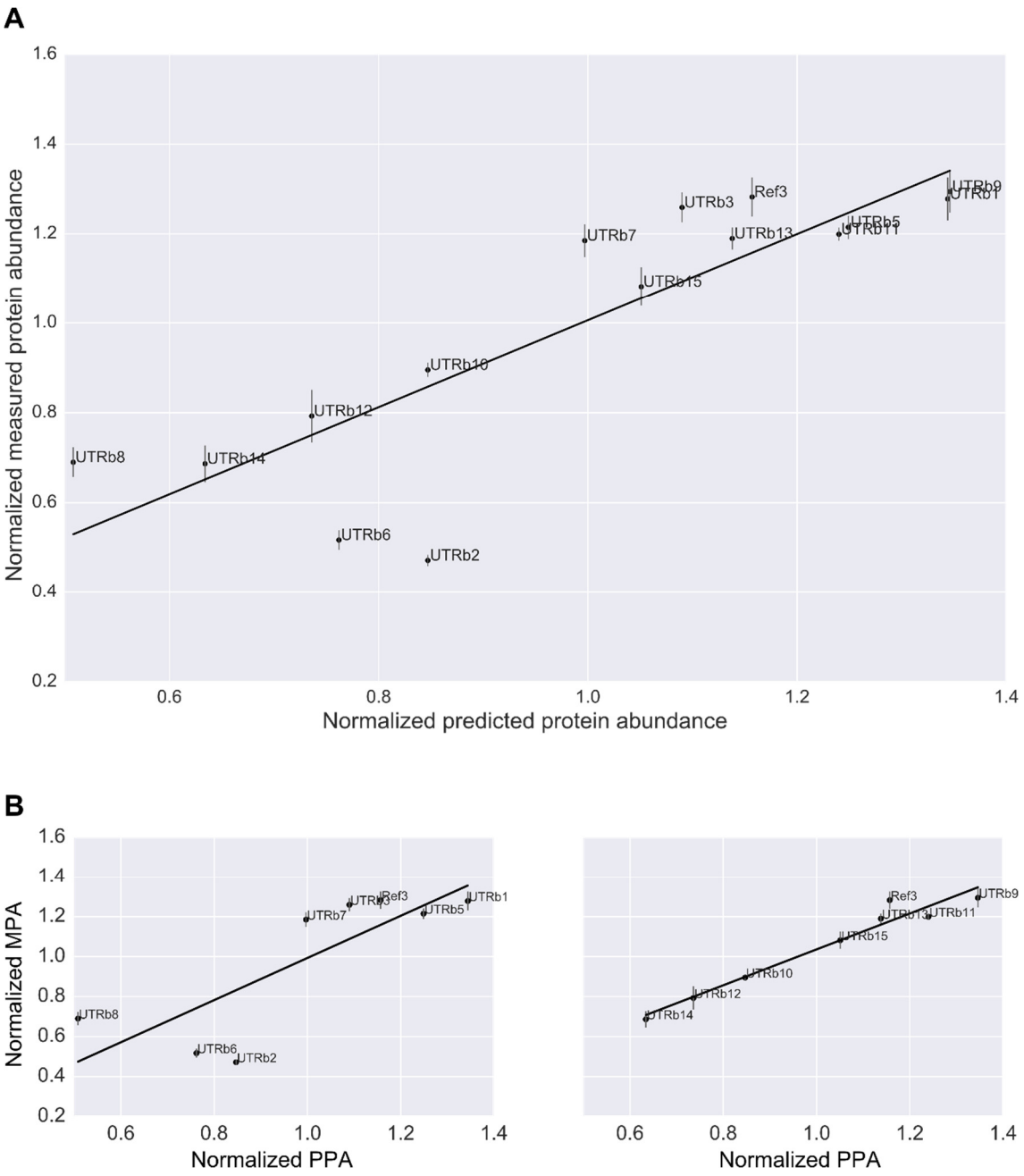

Figure S.2.16: OLS regression plots comparing the normalized predicted protein abundance (PPA) calculated by forward engineering with our model and the normalized measured protein abundance (MPA), determined by measuring mTFP1-to-mCherry ratios. (A) Regression plot of both calculated 8-containing 5'UTR libraries representing strains s_yCIII-1 to s_yCIII-15, additionally, reference strain sTemplate3 was included (R^2 = 0.73). (B) Left: Regression plot of the first part of library UTRb consisting of eight 5'UTR candidates representing strains s_yCIII-1 to s_yCIII-8 including reference strain sTemplate3 (R^2 = 0.65). Right: Regression plot of the second part of library UTRb consisting of eight 5'UTR candidates representing strains s_yCIII-9 to s_yCIII-15 including reference strain sTemplate3 (R^2 = 0.95). Error bars represent standard error of the mean of four biological replicates. Due to cloning issues, strains s_yCIII-4 and s_yCIII-16 are not included.

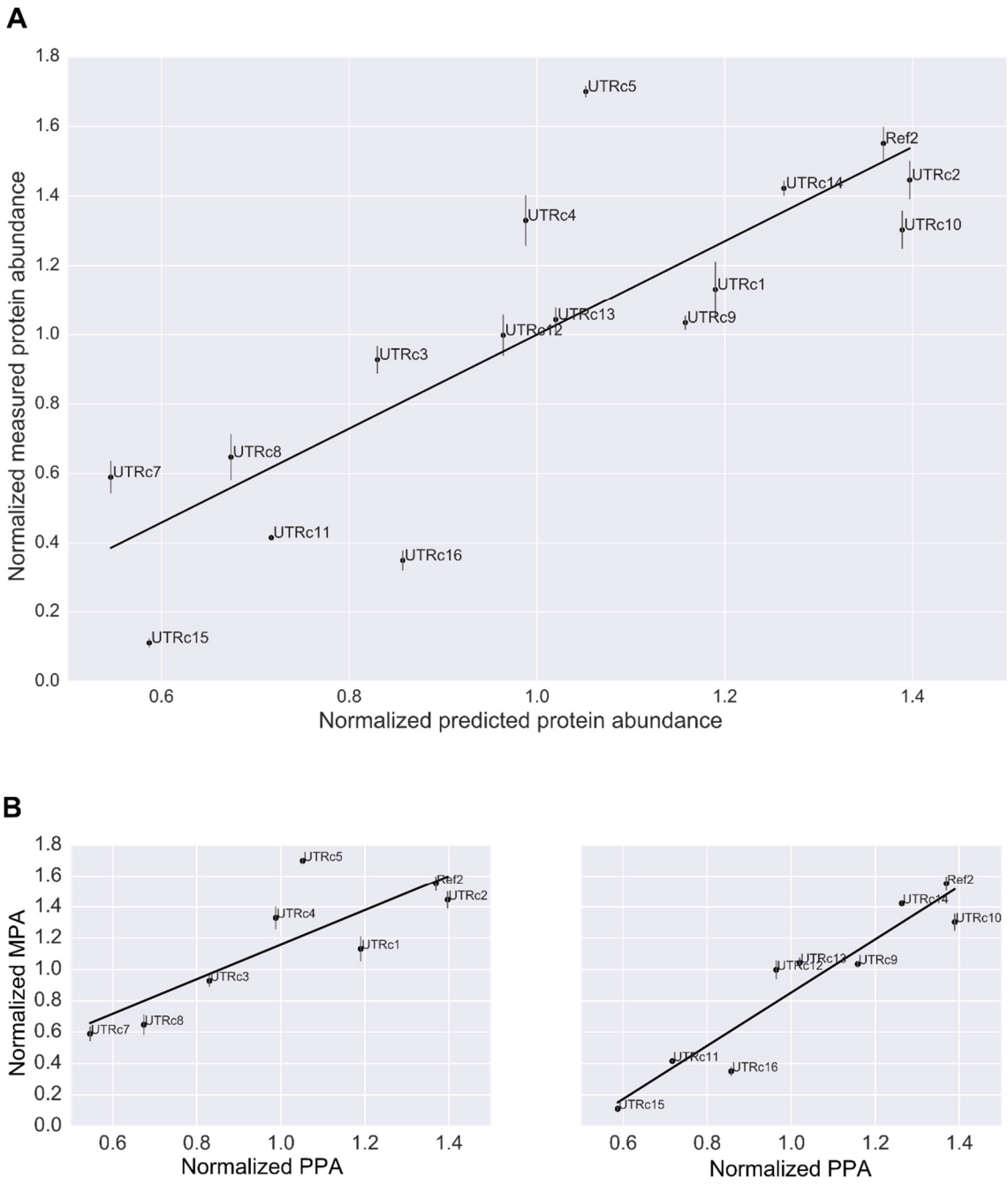

Figure S.2.17: OLS regression plots comparing the normalized predicted protein abundance (PPA) calculated by forward engineering with our model and the normalized measured protein abundance (MPA), determined by measuring yECitrine-to-mCherry ratios. (A) Regression plot of both calculated 8-containing 5'UTR libraries representing strains s_yCIV-1 to s_yCIV-16, additionally, reference strain sTemplate2 was included (R^2 = 0.67). (B) Left: Regression plot of the first part of library UTRc consisting of eight 5'UTR candidates representing strains s_yCIV-1 to s_yCIV-8 including reference strain sTemplate2 (R^2 = 0.69). Right: Regression plot of the second part of library UTRc consisting of eight 5'UTR candidates representing strains s_yCIV-9 to s_yCIV-16 including reference strain sTemplate2 (R^2 = 0.90). Error bars represent standard error of the mean of four biological replicates. Due to cloning issues, strain s_yCIV-6 is not included.

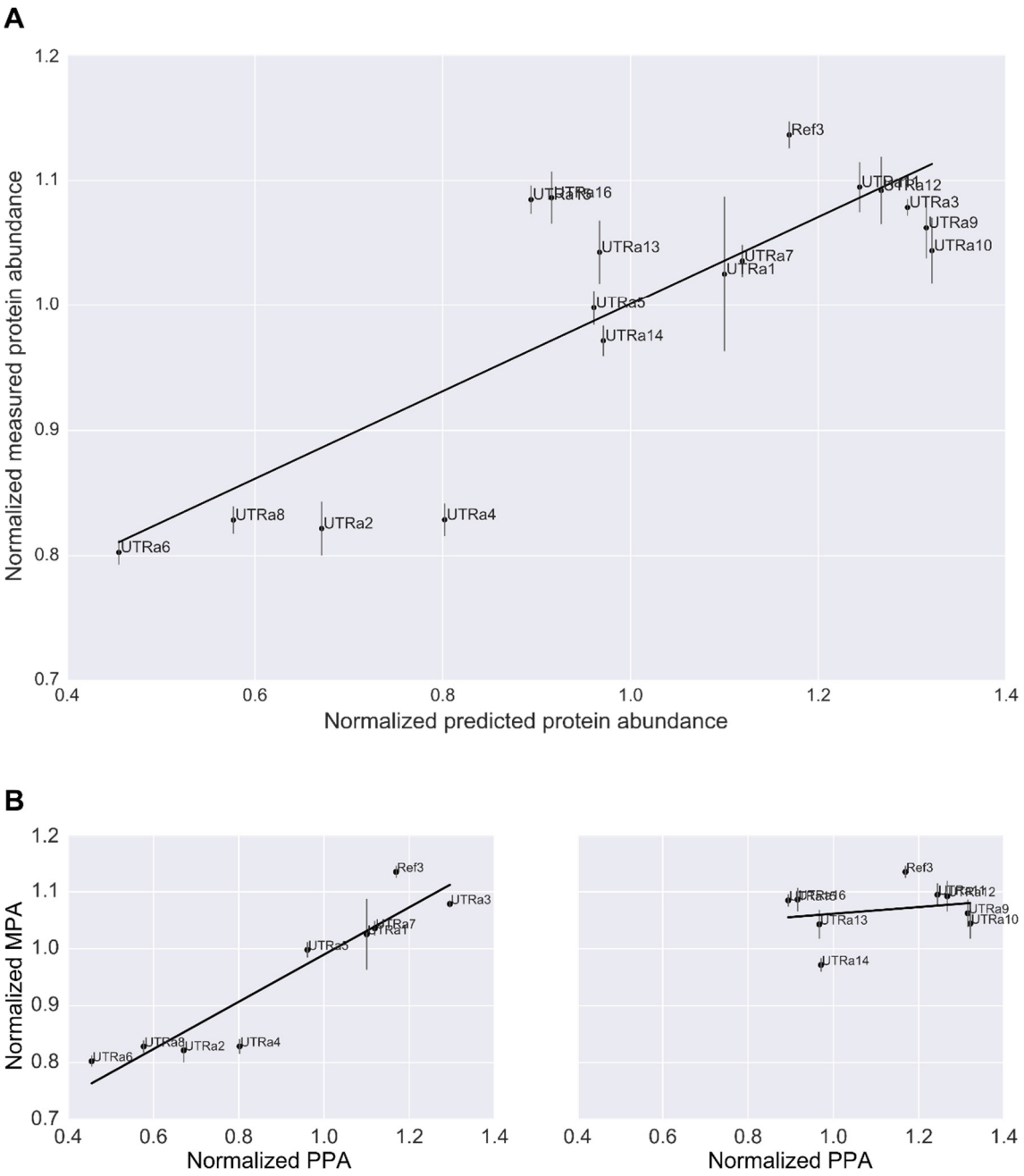

Figure S.2.18: OLS regression plots comparing the normalized predicted protein abundance (PPA) recalculated by reverse engineering with our model and the normalized measured protein abundance (MPA), determined by measuring mTFP1-to-mCherry ratios. (A) Regression plot of both calculated 8-containing 5'UTR libraries representing strains s_yCV-1 to s_yCV-16, additionally, reference strain sTemplate3 was included (R^2 = 0.69). (B) Left: Regression plot of the first part of library UTRa consisting of eight 5'UTR candidates representing strains s_yCV-1 to s_yCV-8 including reference strain sTemplate3 (R^2 = 0.88). Right: Regression plot of the second part of library UTRa consisting of eight 5'UTR candidates representing strains s_yCV-9 to s_yCV-16 including reference strain sTemplate3 (R^2 = 0.05). Error bars represent standard error of the mean of four biological replicates.

S.3 APPENDIX CHAPTER 5

Table S.3.1: Strains used in this study.

Strain	Genotype	Reference
BY4742	*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*	[205]
sRep1	p2a33_yECitrine	This study
sRep2	p2a33_mCherry	This study
sRep3	p2a33_mTFP1	This study
sRep4	p2a33_mTagBFP2	This study
sRepb	p2a_empty	This study
sCas9L	BY4742 + pCas9L	This study
sReg1	sCas9L *ura3Δ0*::pTEF1-yECitrine-tADH1 + p_gRNA_URA3	This study
sReg2	sCas9L *ura3Δ0*::pTEF1-mCherry-tADH1 + p_gRNA_URA3	This study
sReg3	sCas9L *ura3Δ0*::pTEF1-mTFP1-tADH1 + p_gRNA_URA3	This study
sReg4	sCas9L *ura3Δ0*::pTEF1-mTagBFP2-tADH1 + p_gRNA_URA3	This study
sT2A1	pT2A1	This study
sT2A2	pT2A2	This study
sT2A3	pT2A3	This study
sT2A4	pT2Ac1	This study
sT2A5	pT2Ac2	This study
sT2A6	pT2Ap1	This study
sT2A7	pT2Ap2	This study
sT2A8	pT2An1	This study
sT2A9	pT2An2	This study
sT2A10	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A1-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A11	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A2-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A12	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A3-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A13	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2Ac1-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A14	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2Ac2-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A15	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2Ap1-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A16	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2Ap2-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A17	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2An1-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A18	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2An2-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A19	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A1-mTFP1-T2A2-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A20	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A1-mTFP1-T2A3-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A21	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A1-mTFP1-T2Ac1-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A22	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A1-mTFP1-T2Ac2-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A23	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A2-mTFP1-T2A3-mCherry-tADH1 + p_gRNA_URA3	This study

sT2A24	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A2-mTFP1-T2Ac1-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A25	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A2-mTFP1-T2Ac2-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A26	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A3-mTFP1-T2Ac1-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A27	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A3-mTFP1-T2Ac2-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A28	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2Ac1-mTFP1-T2Ac2-mCherry-tADH1 + p_gRNA_URA3	This study
sT2A29	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A1_mTagBFP2_T2A2_mTFP1_ T2A3_mCherry-tADH1 + p_gRNA_URA3	This study
sT2A30	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A1_mTagBFP2_T2A2_mTFP1_ T2Ac1_mCherry-tADH1 + p_gRNA_URA3	This study
sT2A31	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A1_mTagBFP2_T2A2_mTFP1_ T2Ac2_mCherry-tADH1 + p_gRNA_URA3	This study
sT2A32	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A1_mTagBFP2_T2A3_mTFP1_ T2Ac1_mCherry-tADH1 + p_gRNA_URA3	This study
sT2A33	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A1_mTagBFP2_T2A3_mTFP1_ T2Ac2_mCherry-tADH1 + p_gRNA_URA3	This study
sT2A34	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A1_mTagBFP2_T2Ac1_mTFP1_ T2Ac2_mCherry-tADH1 + p_gRNA_URA3	This study
sT2A35	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A2_mTagBFP2_T2A3_mTFP1_ T2Ac1_mCherry-tADH1 + p_gRNA_URA3	This study
sT2A36	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A2_mTagBFP2_T2A3_mTFP1_ T2Ac2_mCherry-tADH1 + p_gRNA_URA3	This study
sT2A37	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A2_mTagBFP2_T2Ac1_mTFP1_ T2Ac2_mCherry-tADH1 + p_gRNA_URA3	This study
sT2A38	sCas9L *ura3Δ0*::pTEF1-yECitrine-T2A3_mTagBFP2_T2Ac1_mTFP1_ T2Ac2_mCherry-tADH1 + p_gRNA_URA3	This study

Table S.3.2: Plasmids used in this study.

Plasmid name	Genotype	Reference
pKT140	yECitrine-tADH1, *KAN*, AmpR, CEN/ARS	[206]
p414	pTEF1-Cas9-tCYC1, *TRP1*, AmpR, CEN/ARS	[19]
p415	pGalL-Cas9-tCYC1, *LEU2*, AmpR, CEN/ARS	[19]
p426	pSNR52-gRNA.CAN1.Y-tSUP4, *URA3,* AmpR, 2µ	[19]
p2a33_yECitrine	pTEF1-yECitrine-tADH1, *URA3*, CEN/ARS	This study
p2a33_mCherry	pTEF1-mCherry-tADH1, *URA3*, CEN/ARS	This study
p2a33_mTFP1	pTEF1-mTFP1-tADH1, *URA3*, CEN/ARS	This study
p2a33_mTagBFP2	pTEF1-mTagBFP2-tADH1, *URA3*, CEN/ARS	This study
p2a_empty	*URA3*, CEN/ARS	This study
pT2A1	pTEF1-yECitrine-T2A1-mCherry-tADH1, *URA3*, CEN/ARS	This study
pT2A2	pTEF1-yECitrine-T2A2-mCherry-tADH1, *URA3*, CEN/ARS	This study
pT2A3	pTEF1-yECitrine-T2A3-mCherry-tADH1, *URA3*, CEN/ARS	This study
pT2Ac1	pTEF1-yECitrine-T2Ac1-mCherry-tADH1, *URA3*, CEN/ARS	This study
pT2Ac2	pTEF1-yECitrine-T2Ac2-mCherry-tADH1, *URA3*, CEN/ARS	This study
pT2Ap1	pTEF1-yECitrine-T2Ap1-mCherry-tADH1, *URA3*, CEN/ARS	This study
pT2Ap2	pTEF1-yECitrine-T2Ap2-mCherry-tADH1, *URA3*, CEN/ARS	This study
pT2An1	pTEF1-yECitrine-T2An1-mCherry-tADH1, *URA3*, CEN/ARS	This study
pT2An2	pTEF1-yECitrine-T2An2-mCherry-tADH1, *URA3*, CEN/ARS	This study
pCas9L	pTEF1-Cas9-tCYC1, *LEU2*, AmpR, CEN/ARS	This study
p_gRNA_URA3	pSNR52-gRNA_URA3-tSUP4, *URA3,* AmpR, 2µ	This study

Table S.3.3: Theoretical possibilities of protein products on Western blot after T2A mediated splicing. All protein sizes are expressed in kDa. Abbreviations: TU: Transcription Unit; yECit: yECitrine; mCh: mCherry.

Splicing	Protein products	Anti-GFP	Anti-mCherry	Anti-2A
TU: yECit-T2A$_a$-mCh				
T2A$_a$	yECit-T2A$_a$	28.6	-	28.6
	mCh	-	26.3	-
None	yECit-T2A$_a$-mCh	54.9	54.9	54.9
TU: yECit-T2A$_a$-mTFP1-T2A$_b$-mCh				
T2A$_a$ & T2A$_b$	yECit-T2A$_a$	-	-	28.6
	mTFP1-T2A$_b$	-	-	28.6
	mCh	-	26.3	-
T2A$_a$	yECit-T2A$_a$	-	-	28.6
	mTFP1-T2A$_b$-mCh	-	54.9	54.9
T2A$_b$	yECit-T2A$_a$-mTFP1-T2A$_b$	-	-	57.2
	mCh	-	26.3	-
None	yECit-T2A$_a$-mTFP1-T2A$_b$-mCh	-	83.5	83.5
TU: yECit-T2A$_a$-mTagBFP2-T2A$_b$-mTFP1-T2A$_c$-mCh				
T2A$_a$, T2A$_b$ & T2A$_c$	yECit-T2A$_a$	-	-	28.6
	mTagBFP2-T2A$_b$	-	-	28.1
	mTFP1-T2A$_c$	-	-	28.6
	mCh	-	26.3	-
T2A$_a$ & T2A$_b$	yECit-T2A$_a$	-	-	28.6
	mTagBFP2-T2A$_b$	-	-	28.1
	mTFP1-T2A$_c$-mCh	-	54.4	54.6
T2A$_a$ & T2A$_c$	yECit-T2A$_a$	-	-	28.6
	mTagBFP2-T2A$_b$-mTFP1-T2A$_c$	-	-	56.7
	mCh	-	26.3	-
T2A$_b$ & T2A$_c$	yECit-T2A$_a$-mTagBFP2-T2A$_b$	-	-	56.7
	mTFP1-T2A$_c$	-	-	28.6
	mCh	-	26.3	-
T2A$_a$	yECit-T2A$_a$	-	-	28.6
	mTagBFP2-T2A$_b$-mTFP1-T2A$_c$-mCh	-	82.7	82.7
T2A$_b$	yECit-T2A$_a$-mTagBFP2-T2A$_b$	-	-	56.7
	mTFP1-T2A$_c$-mCh	-	54.9	54.9
T2A$_c$	yECit-T2A$_a$-mTagBFP2-T2A$_b$-mTFP1-T2A$_c$	-	-	85.3
	mCh	-	26.3	-
None	yECit-T2A$_a$-mTagBFP2-T2A$_b$-mTFP1-T2A$_c$-mCh	-	111.6	111.6

A) Multiple sequence alignment of Geier's modified T2A sequences with GSG-tag

```
T2A1*      GGTTCTGGTGAGGGTAGAGGTTCTTTGCTTACTTGCGGTGACGTTGAGGAAAACCCAGGACCT 63
T2A2*      GGTTCAGGAGAAGGACGTGGATCCCTTTTGACCTGCGGAGATGTCGAAGAGAATCCTGGACCT 63
T2A3*      GGTTCCGGCGAAGGTCGTGGCTCATTGCTGACTTGTGGCGACGTGGAGGAAAATCCCGGACCT 63
T2A4*      GGTTCGGGGGAGGGCCGTGGTTCCTTACTTACCTGCGGTGATGTGGAAGAAAATCCAGGACCT 63
T2A5*      GGTTCGGGGGAGGGTAGGGGATCACTTCTTACATGTGGAGACGTCGAGGAGAACCCTGGACCT 63
T2A6*      GGTTCCGGCGAAGGAAGGGGTTCCCTGTTAACGTGTGGCGATGTTGAAGAGAACCCCGGACCT 63
T2A7*      GGTTCAGGAGAAGGCAGAGGATCTCTGTTGACTTGTGGTGATGTAGAGGAGAATCCCGGACCT 63
T2A8*      GGTTCTGGTGAGGGGAGAGGCTCTCTTTTAACTTGTGGAGATGTGGAAGAGAACCCAGGACCT 63
           ***** ** ** **  * ** **  *  * ** ** ** ** ** ** ** ** ** ******
```

```
Percent Identity  Matrix
```

T2A1*	100.00	68.25	79.37	80.95	79.37	74.60	77.78	80.95
T2A2*	68.25	100.00	74.60	79.37	77.78	79.37	82.54	77.78
T2A3*	79.37	74.60	100.00	77.78	76.19	77.78	79.37	73.02
T2A4*	80.95	79.37	77.78	100.00	74.60	73.02	73.02	74.60
T2A5*	79.37	77.78	76.19	74.60	100.00	76.19	76.19	79.37
T2A6*	74.60	79.37	77.78	73.02	76.19	100.00	80.95	80.95
T2A7*	77.78	82.54	79.37	73.02	76.19	80.95	100.00	80.95
T2A8*	80.95	77.78	73.02	74.60	79.37	80.95	80.95	100.00

B) Multiple sequence alignment of T2A sequences derived from Geier's sequences after the introduction of mutations

```
T2A1       GGTTCTGGTGAAGGTAGAGGTTCTTTGCTTACTTGCGGTGATGTTGAGGAAAACCCAGGACCT 63
T2A2       GGTTCAGGAGAAGGACGTGGAAGCCTTTTGACCTGCGGAGATGTCGAAGAGAATCCTGGACCT 63
T2A3       GGTTCCGGCGAGGGCAGGGGCTCACTGTTAACGTGTGGCGACGTGGAATCTAACCCCGGACCT 63
           ***** ** ** **  * **     *  * ** ** ** ** ** **    ** ** ******
```

```
Percent Identity  Matrix
```

T2A1	100.00	68.25	68.25
T2A2	68.25	100.00	65.08
T2A3	68.25	65.08	100.00

C) Multiple sequence alignment of the final five new designed T2A sequences

```
T2A1       GGTTCTGGTGAAGGTAGAGGTTCTTTGCTTACTTGCGGTGATGT---TGAGGAAAACCCAGGACCT 63
T2A2       GGTTCAGGAGAAGGACGTGGAAGCCTTTTGACCTGCGGAGATGT---CGAAGAGAATCCTGGACCT 63
T2A3       GGTTCCGGCGAGGGCAGGGGCTCACTGTTAACGTGTGGC---GACGTGGAATCTAACCCCGGACCT 63
T2Ac1      GGTTCTGGTGAAGGAAGGGGTTCTTTGTTGACTCTAGCAGGTGACGTCGAATCTAACCCTGGACCT 66
T2Ac2      GGTTCAGGATTGCTTAATTTTGATCTTCTTAAGCTTTGTGGAGATGTTGAGGAGAATCCAGGACCT 66
           ***** **                *  * *           *     **    ** ** ******
```

```
Percent Identity  Matrix
```

T2A1	100.00	68.25	66.67	73.02	58.73
T2A2	68.25	100.00	63.33	66.67	53.97
T2A3	66.67	63.33	100.00	74.60	50.79
T2Ac1	73.02	66.67	74.60	100.00	50.00
T2Ac2	58.73	53.97	50.79	50.00	100.00

Figure S.3.1: Results of Clustal Omega sequence alignments of different T2A peptide sequences. (A) Sequence alignment and percent identity matrix of the T2A sequences obtained from Geier *et al.* [286] after introduction of GSG-tags consisting of different codons and equalizing the four nucleotides at the 5' and 3' end for Golden Gate purposes. (B) Sequence alignment and identity matrix of the first three T2A sequences designed for this study. Sequences were derived from Geier *et al.* [286] after introduction of different silent mutations and a codon replacement (S17E) in T2A6*. (C) Sequence alignment and identity matrix of the five sequence optimized T2A sequences further used in this study.

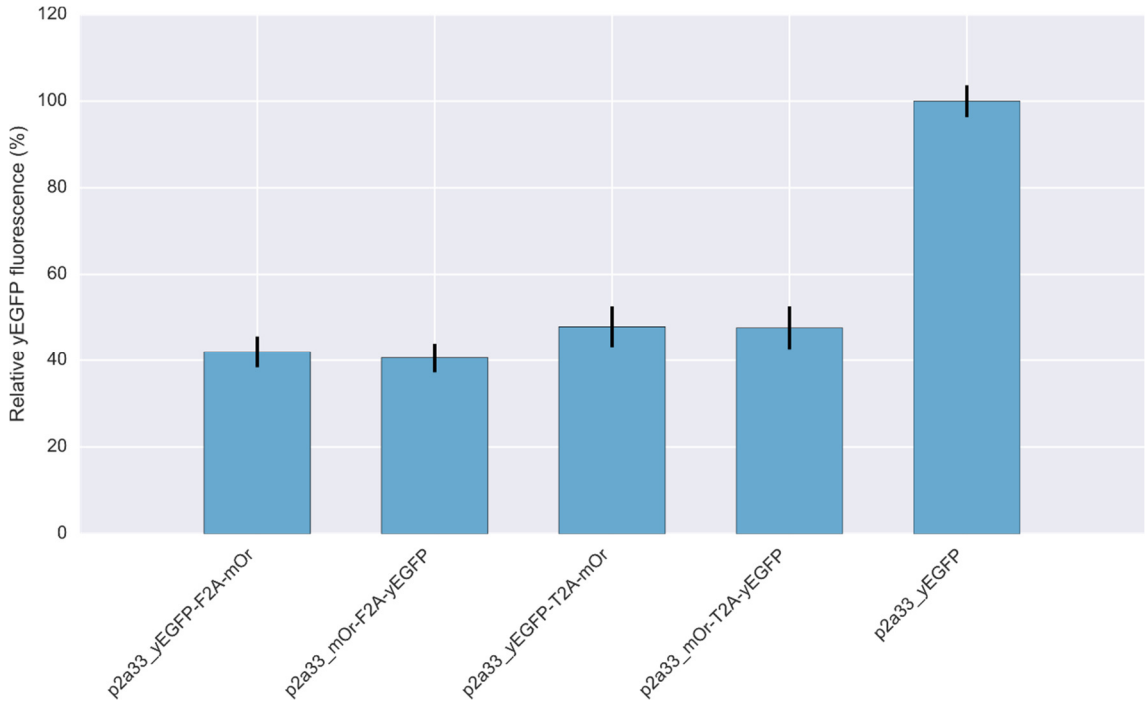

Figure S.3.2: Influence of the protein coding sequence position on fluorescence. Fluorescence was normalized against the monocistronic reference p2a33_yEGFP. Error bars represent the standard error of the mean of four biological replicates. Abbreviations: F2A: 2A peptide originating from foot-and-mouth disease virus; T2A, 2A peptide originating from the *Thosea asigna* virus.

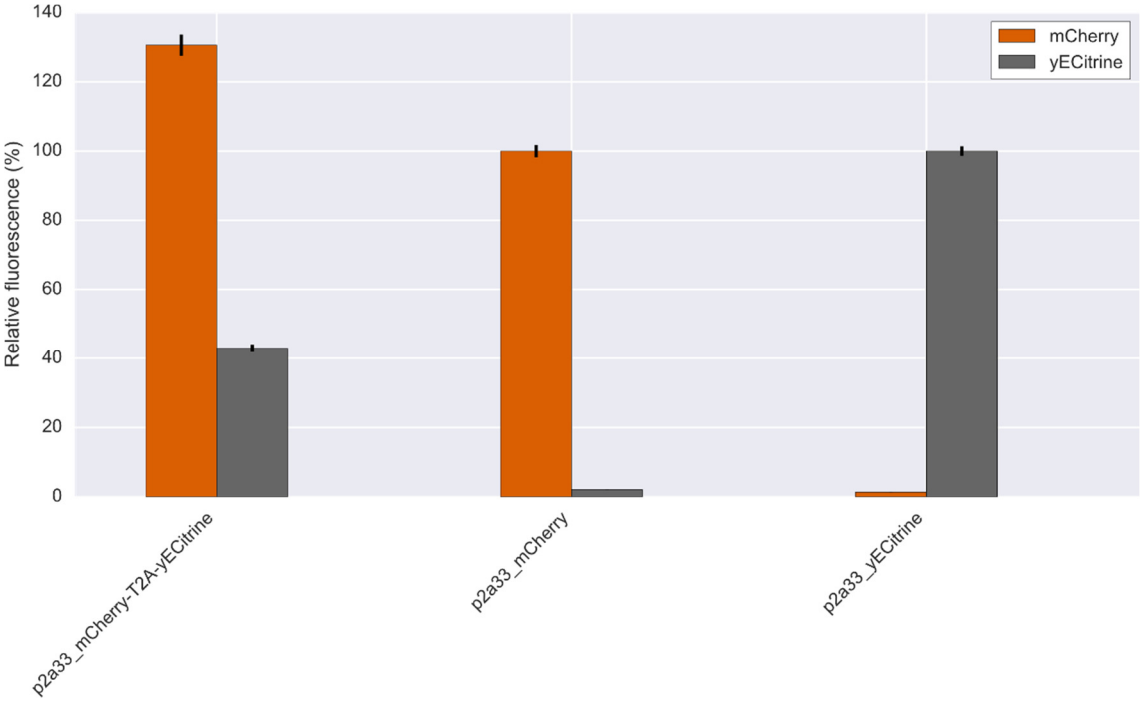

Figure S.3.3: Evaluation of the positioning effect on fluorescence with mCherry and yECitrine. Fluorescence of both fluorescent reporters was normalized to their monocistronic reference strains p2a33_mCherry and p2a33_yECitrine. Error bars represent the standard error of the mean of four biological replicates. Abbreviations: T2A, 2A peptide originating from the *Thosea asigna* virus.

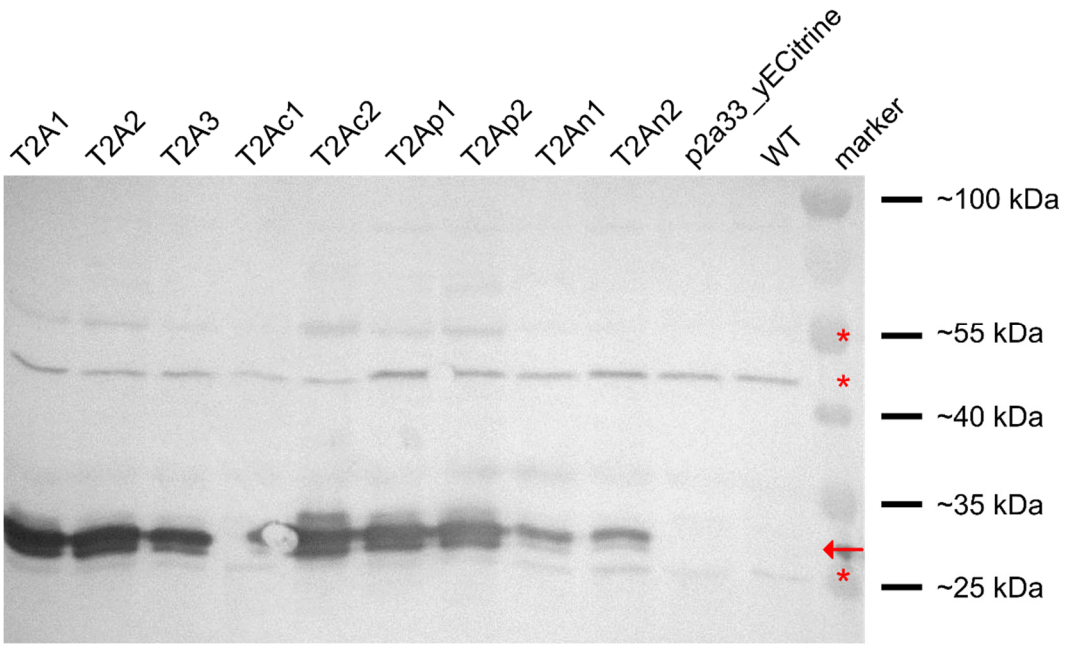

Figure S.3.4: Characterization of splicing efficiency of the T2A peptides expressed on a low copy, bicistronic expression vector using Western blot with anti-2A antiserum. The red arrow represents the size of cleaved proteins and asterisks indicate unknown detected byproducts. The strains corresponding with the represented T2A peptides are listed in Supplementary Table S.3.1.

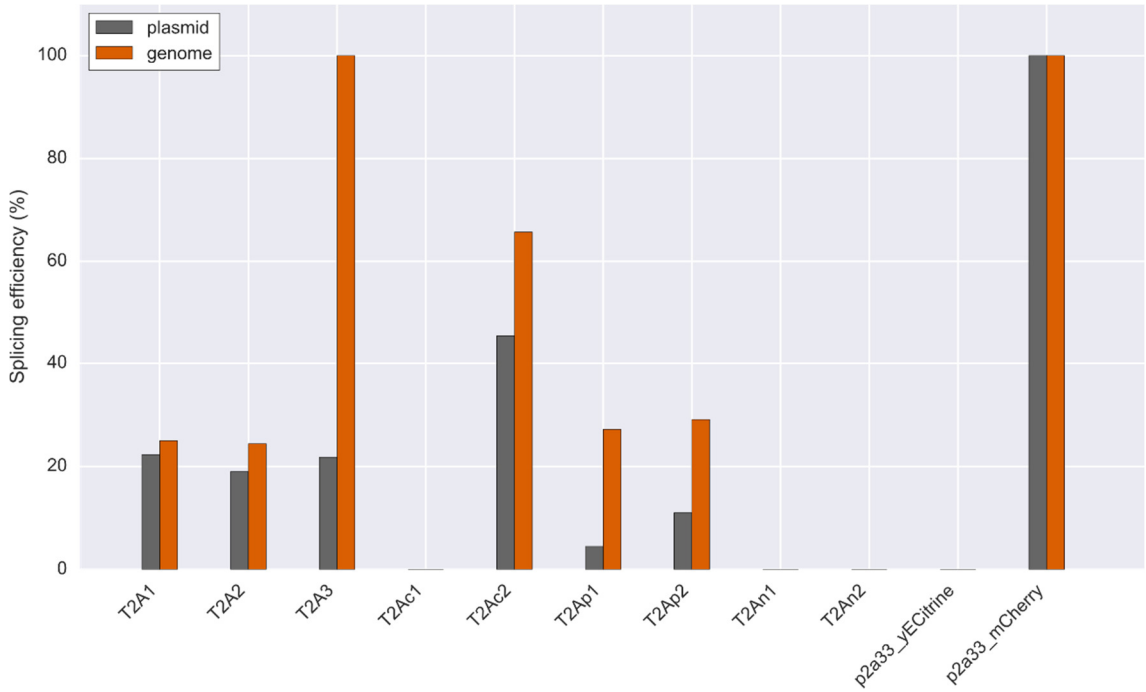

Figure S.3.5: Quantification of splicing efficiencies of the indicated T2A sequences determined by ImageJ software. Splicing efficiency was calculated as follows: spliced protein/(spliced protein + unspliced protein). Efficiencies were determined with the anti-mCherry Western blots of the pTEF1-yECit-T2A$_a$-mCh-tADH1 transcription units expressed on plasmid and genome, respectively. Note: For T2A3 on the genome, the splicing efficiency is anomalous since protein amounts loaded on SDS gel were too low.

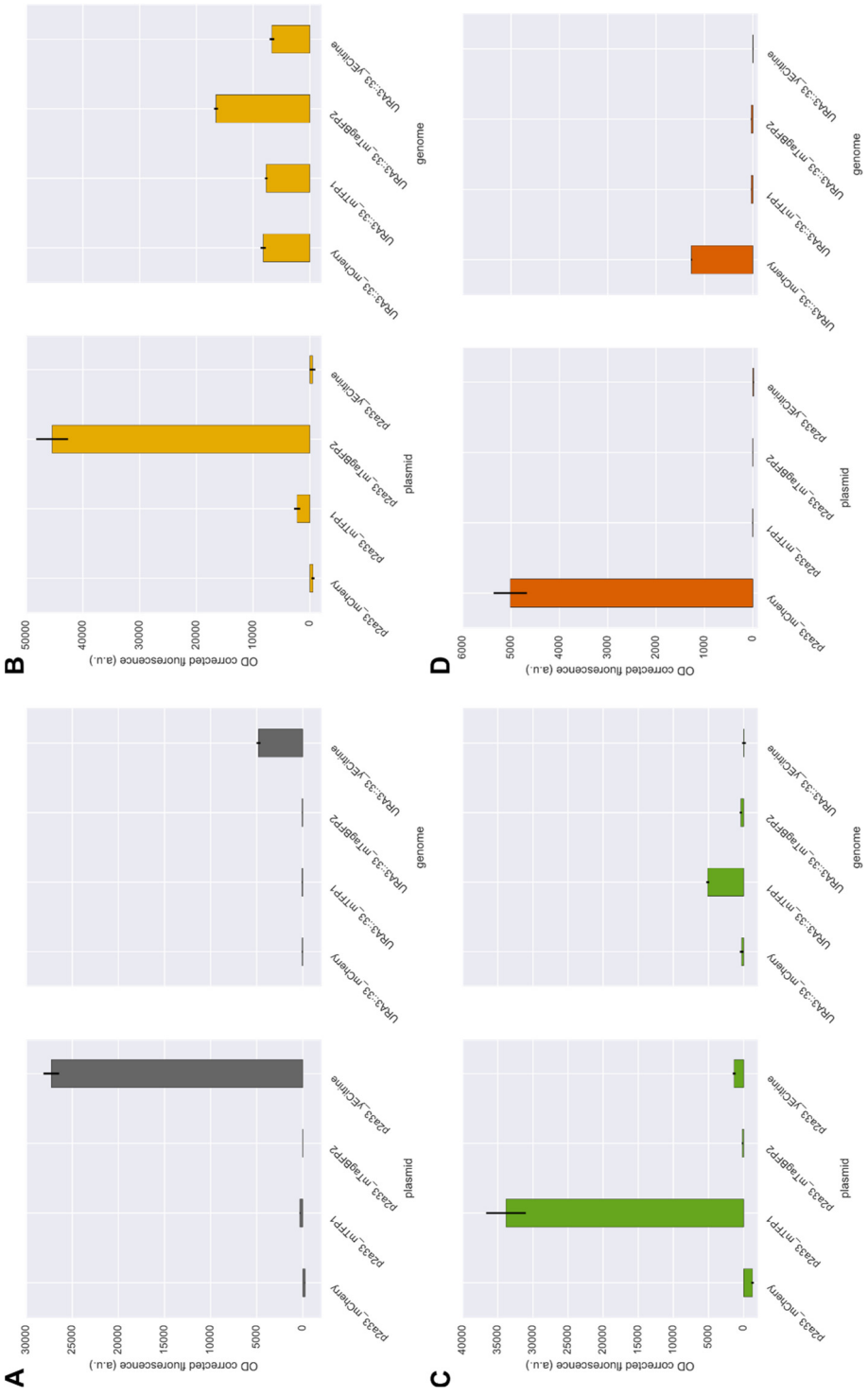

Figure S.3.6: Characterization of fluorescent proteins either expressed from a low-copy expression backbone (plasmid, CEN6/ARS4, *URA3*) or expressed from the *URA3* locus (genome) in *S. cerevisiae* BY4742. (A) yECitrine fluorescence (ex. 500 nm, em. 540 nm). (B) mTagBFP2 fluorescence (ex. 415 nm, em. 460 nm). (C) mTFP1 fluorescence (ex. 460 nm, em. 500 nm). (D) mCherry fluorescence (ex. 575 nm, em. 620 nm). All measurements were run in a TECAN Infinite® 200 PRO (Tecan) plate reader. Error bars represent the standard error of the mean of three biological replicates.

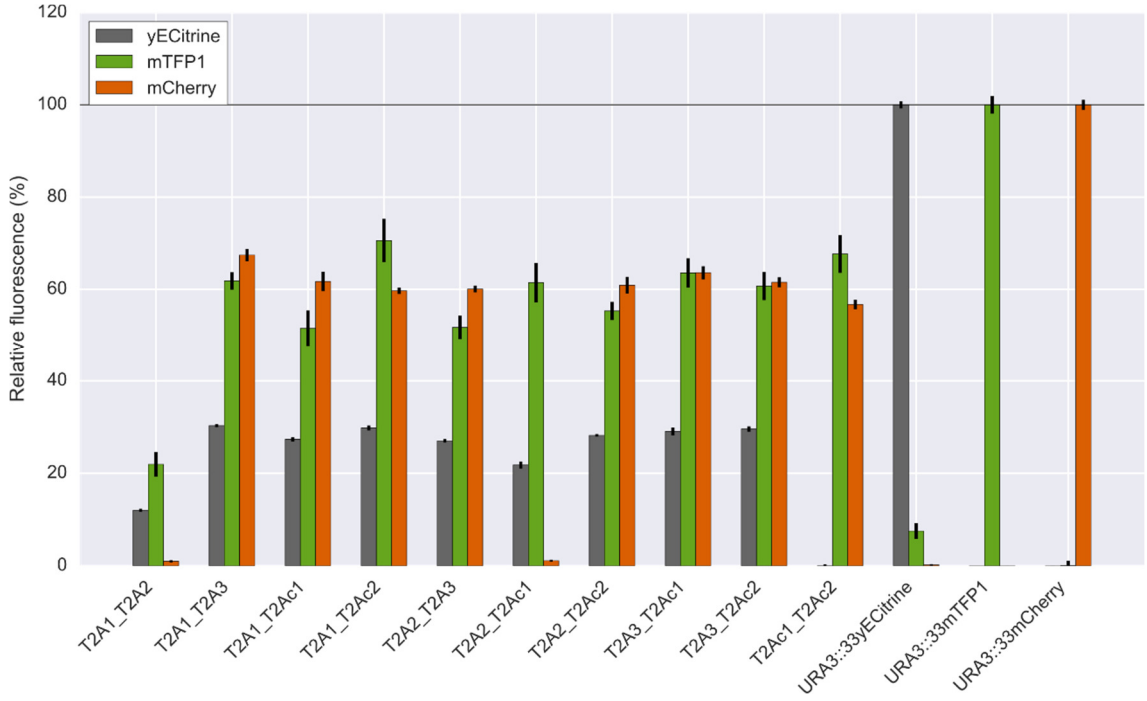

Figure S.3.7: Gene expression analysis using fluorescence measurements of yECitrine, mTFP1 and mCherry in a tricistronic construct normalized to their monocistronic reference strains (represented by the horizontal line). Shown here are the results of a second, independent fluorescence measurement to support data in Figure 5.5. Error bars represent the standard error of the mean of three biological replicates.

S.4 APPENDIX CHAPTER 6

Table S.4.1: Nucleotide sequences of the codon harmonized flavonoid genes used in this study. Abbreviations: At: *Arabidopsis thaliana,* Rc: *Rhodobacter capsulatus,* Gm: *Glycine max*.

AtPAL1

ATGGAGATTAACGGGGCACACAAAAGCAACGGTGGTGGTGTAGACGCTATGCTATGCGG
GGGTGACATCAAAACAAAAAACATGGTCATCAACGCGGAGGATCCACTGAACTGGGGTG
CTGCAGCGGAGCAAATGAAGGGAAGCCATTTAGATGAAGTAAAAAGAATGGTTGCTGAG
TTCAGGAAACCTGTTGTCAACTTAGGAGGAGAGACTCTGACCATTGGTCAAGTAGCTGC
GATATCAACTATTGGAAACAGTGTAAAAGTAGAGCTGAGCGAGACAGCTAGAGCGGGAG
TAAATGCTAGTAGTGATTGGGTTATGGAGAGTATGAACAAGGGGACTGATAGTTATGGA
GTTACTACTGGATTTGGAGCTACTTCTCATCGTAGAACCAAGAACGGAGTGGCCTTGCA
GAAGGAGTTGATTAGATTCTTGAACGCGGGTATCTTCGGTAGCACGAAGGAAACAAGCC
ACACATTACCTCACAGCGCGACAAGAGCGGCGATGTTAGTGCGTATAAACACTCTACTA
CAAGGTTTTAGCGGAATACGTTTTGAGATTCTAGAAGCAATTACCAGTTTCCTAAACAA
CAACATAACTCCTTCTCTACCGCTACGTGGAACAATAACCGCGAGCGGTGATCTAGTTC
CACTAAGCTACATAGCGGGTTTACTAACCGGACGTCCGAATAGCAAGGCTACTGGACCG
AACGGAGAAGCTCTAACAGCAGAGGAAGCTTTCAAGCTAGCAGGTATAAGCAGCGGTTT
CTTTGATCTACAGCCCAAGGAAGGACTAGCGCTGGTGAATGGGACGGCGGTTGGTTCTG
GTATGGCGTCAATGGTACTATTCGAAACGAATGTTCTATCTGTTTTAGCTGAGATTTTA
AGCGCGGTTTTCGCAGAGGTAATGAGTGGAAAACCAGAGTTCACCGATCATCTAACTCA
CAGATTAAAGCATCATCCGGGACAAATAGAAGCGGCGGCGATCATGGAGCATATACTAG
ACGGTAGCAGCTACATGAAGCTAGCTCAGAAATTACACGAGATGGATCCCCTACAGAAG
CCAAAGCAAGATCGTTACGCTTTACGTACTTCTCCACAATGGCTAGGACCACAAATAGA
AGTAATACGTTACGCAACGAAGAGCATAGAGCGTGAGATTAACAGCGTGAACGATAATC
CCTTAATAGATGTTTCGAGGAACAAAGCGATTCACGGAGGAAACTTCCAAGGTACACCT
ATAGGTGTTTCAATGGATAACACGAGATTAGCGATCGCAGCGATTGGAAAGCTAATGTT
TGCTCAATTCTCAGAGTTAGTAAATGATTTCTACAACAATGGACTACCGAGCAATCTGA
CCGCTTCGAGGAATCCTAGTTTAGATTATGGTTTCAAAGGTGCTGAGATTGCAATGGCT
TCATATTGTTCAGAGTTACAATACCTAGCTAATCCAGTAACTAGCCATGTTCAATCAGC
AGAGCAACATAACCAAGATGTGAACTCTTTAGGTCTGATAAGCTCTCGAAAGACTTCTG
AAGCTGTTGATATTCTAAAATTAATGTCAACAACGTTCCTAGTTGCGATTTGTCAAGCT
GTAGATTTAAGACATTTAGAGGAGAATTTAAGACAGACTGTAAAAAACACTGTGTCTCA
AGTAGCGAAAAAGGTCCTAACTACTGGTGTGAATGGAGAGTTACATCCATCTCGATTCT
GCGAAAAGGATCTACTAAAGGTTGTGGACCGTGAACAAGTGTACACATACGCGGATGAT
CCATGTAGCGCAACGTACCCCTTAATTCAGAAACTGAGACAAGTTATTGTTGACCATGC
TTTAATCAATGGAGAGAGTGAGAAAAATGCAGTAACTTCTATATTCCACAAGATAGGTG
CTTTCGAGGAGGAGTTAAAAGCAGTACTGCCCAAGGAAGTAGAAGCAGCAAGAGCAGCG
TACGATAACGGTACATCGGCTATACCCAACAGGATAAAGGAATGTAGGAGCTATCCTTT
ATATAGATTCGTAAGGGAAGAGTTAGGTACAGAGTTATTAACCGGTGAGAAGGTAACGT
CGCCAGGTGAAGAGTTCGACAAAGTTTTCACGGCGATTTGTGAAGGAAAGATAATTGAT
CCCATGATGGAATGTCTAAACGAGTGGAACGGTGCTCCGATTCCTATCTGTTGA

AtC4H

ATGGACCTACTATTACTGGAGAAATCTCTAATAGCGGTGTTCGTAGCGGTAATTCTAGC
GACGGTAATTTCAAAACTACGAGGGAAAAAGTTGAAACTGCCACCTGGACCTATCCCTA
TTCCCATATTCGGTAACTGGTTACAAGTGGGTGATGATCTAAACCACCGTAATCTAGTG
GATTACGCTAAAAAGTTCGGGGATCTATTCCTACTACGTATGGGACAGCGTAACCTGGT
GGTGGTGAGCTCACCCGATCTGACAAAGGAAGTACTACTAACTCAAGGGGTTGAGTTTG
GTAGCAGAACGAGAAACGTGGTATTCGACATTTTCACCGGGAAGGGACAAGATATGGTA
TTCACTGTTTACGGGGAGCATTGGAGGAAAATGAGAAGAATCATGACGGTTCCATTCTT
CACCAACAAGGTTGTTCAACAGAATCGTGAAGGATGGGAGTTTGAAGCAGCTAGTGTTG
TTGAAGATGTTAAAAAAAATCCTGATTCTGCTACGAAGGGTATAGTCTTGAGGAAGCGT
TTACAATTAATGATGTATAACAATATGTTCCGTATAATGTTCGATAGAAGATTTGAGAG
TGAGGATGATCCATTATTCCTAAGGTTAAAAGCTTTAAATGGAGAGAGAAGTCGTCTAG

CTCAGAGCTTTGAGTATAACTATGGTGATTTCATTCCAATATTAAGACCTTTCCTAAGA
GGGTATTTGAAAATTTGTCAAGATGTAAAGGATCGTAGAATAGCATTATTCAAAAAATA
CTTCGTTGATGAGAGGAAACAGATAGCGAGTTCTAAACCAACAGGAAGTGAAGGTTTAA
AGTGTGCGATTGATCACATATTAGAAGCTGAGCAGAAAGGTGAAATCAACGAGGACAAT
GTTTTATACATAGTGGAGAACATAAATGTGGCGGCGATTGAGACAACATTATGGTCTAT
AGAGTGGGGTATTGCAGAGCTGGTAAACCATCCAGAAATACAGAGTAAACTGAGGAACG
AACTAGACACAGTTTTAGGTCCCGGAGTACAAGTGACCGAGCCAGATTTACACAAGTTA
CCTTACTTACAAGCTGTAGTTAAAGAGACTTTACGTCTGAGAATGGCGATTCCACTACT
AGTACCACACATGAACCTACATGATGCGAAACTAGCTGGGTACGATATACCTGCAGAAA
GCAAGATATTAGTTAATGCTTGGTGGCTGGCAAACAACCCGAACAGCTGGAAAAAACCA
GAAGAGTTCAGACCTGAGAGGTTCTTCGAAGAAGAATCGCACGTAGAAGCTAACGGAAA
TGACTTCAGGTATGTACCTTTTGGAGTTGGTCGTCGTAGCTGTCCGGGGATTATCTTAG
CATTACCAATTTTAGGGATAACCATTGGAAGGATGGTGCAGAACTTCGAGTTATTACCA
CCACCTGGTCAGTCTAAGGTAGATACTAGTGAGAAGGGAGGTCAATTCAGCTTACACAT
ATTAAACCACAGCATCATAGTTATGAAGCCTAGGAACTGTTGA

AtATR1 ATGACTTCTGCTTTGTATGCTTCCGATTTGTTTAAGCAGCTCAAGTCAATTATGGGGAC
AGATTCGTTATCCGACGATGTTGTACTTGTGATTGCAACGACGTCTTTGGCACTAGTAG
CTGGATTTGTGGTGTTGTTATGGAAGAAAACGACGGCGGATCGGAGCGGGGAGCTGAAG
CCTTTGATGATCCCTAAGTCTCTTATGGCTAAGGACGAGGATGATGATTTGGATTTGGG
ATCCGGGAAGACTAGAGTCTCTATCTTCTTCGGTACGCAGACTGGAACAGCTGAGGGAT
TTGCTAAGGCATTATCCGAAGAAATCAAAGCGAGATATGAAAAAGCAGCAGTCAAAGTC
ATTGACTTGGATGACTATGCTGCCGATGATGACCAGTATGAAGAGAAATTGAAGAAGGA
AACTTTGGCATTTTTCTGTGTTGCTACTTATGGAGATGGAGAGCCTACTGACAATGCTG
CCAGATTTTACAAATGGTTTACGGAGGAAAATGAACGGGATATAAAGCTTCAACAACTA
GCATATGGTGTGTTTGCTCTTGGTAATCGCCAATATGAACATTTTAATAAGATCGGGAT
AGTTCTTGATGAAGAGTTATGTAAGAAAGGTGCAAAGCGTCTTATTGAAGTCGGTCTAG
GAGATGATGATCAGAGCATTGAGGATGATTTTAATGCCTGGAAAGAATCACTATGGTCT
GAGCTAGACAAGCTCCTCAAAGACGAGGATGATAAAAGTGTGGCAACTCCTTATACAGC
TGTTATTCCTGAATACCGGGTGGTGACTCATGATCCTCGGTTTACAACTCAAAAATCAA
TGGAATCAAATGTGGCCAATGGAAATACTACTATTGACATTCATCATCCCTGCAGAGTT
GATGTTGCTGTGCAGAAGGAGCTTCACACACATGAATCTGATCGGTCTTGCATTCATCT
CGAGTTCGACATATCCAGGACGGGTATTACATATGAAACAGGTGACCATGTAGGTGTAT
ATGCTGAAAATCATGTTGAAATAGTTGAAGAAGCTGGAAAATTGCTTGGCCACTCTTTA
GATTTAGTATTTTCCATACATGCTGACAAGGAAGATGGCTCCCCATTGGAAAGCGCAGT
GCCGCCTCCTTTCCCTGGTCCATGCACACTTGGGACTGGTTTGGCAAGATACGCAGACC
TTTTGAACCCTCCTCGAAAGTCTGCGTTAGTTGCCTTGGCGGCCTATGCCACTGAACCA
AGTGAAGCCGAGAAACTTAAGCACCTGACATCACCTGATGGAAAGGATGAGTACTCACA
ATGGATTGTTGCAAGTCAGAGAAGTCTTTTAGAGGTGATGGCTGCTTTTCCATCTGCAA
AACCCCCACTAGGTGTATTTTTTGCTGCAATAGCTCCTCGTCTACAACCTCGTTACTAC
TCCATCTCATCCTCGCCAAGATTGGCGCCAAGTAGAGTTCATGTTACATCCGCACTAGT
ATATGGTCCAACTCCTACTGGTAGAATCCACAAGGGTGTGTGTTCTACGTGGATGAAGA
ATGCAGTTCCTGCGGAGAAAAGTCATGAATGTAGTGGAGCCCCAATCTTTATTCGAGCA
TCTAATTTCAAGTTACCATCCAACCCTTCAACTCCAATCGTTATGGTGGGACCTGGGAC
TGGGCTGGCACCTTTTAGAGGTTTTCTGCAGGAAAGGATGGCACTAAAAGAAGATGGAG
AAGAACTAGGTTCATCTTTGCTCTTCTTTGGGTGTAGAAATCGACAGATGGACTTTATA
TACGAGGATGAGCTCAATAATTTTGTTGATCAAGGCGTAATATCTGAGCTCATCATGGC
ATTCTCCCGTGAAGGAGCTCAGAAGGAGTATGTTCAACATAAGATGATGGAGAAGGCAG
CACAAGTTTGGGATCTAATAAAGGAAGAAGGATATCTCTATGTATGCGGTGATGCTAAG
GGCATGGCGAGGGACGTCCACCGAACTCTACACACCATTGTTCAGGAGCAGGAAGGTGT
GAGTTCGTCAGAGGCAGAGGCTATAGTTAAGAAACTTCAAACCGAAGGAAGATACCTCA
GAGATGTCTGGTGA

RcTal1

ATGACCTTACAATCCCAAACTGCCAAAGACTGCTTAGCCTTAGACGGTGCCTTGACCTT
GGTTCAATGTGAAGCAATTGCCACACATAGATCCAGAATAAGTGTCACCCCAGCTTTGA
GAGAAAGATGCGCTAGAGCACATGCCAGATTAGAACACGCTATTGCAGAACAAAGACAC
ATCTATGGTATAACTACAGGTTTTGGTCCTTTGGCTAATAGATTAATAGGTGCCGATCA
AGGTGCTGAATTGCAACAAAACTTAATCTACCATTTGGCTACTGGTGTTGGTCCAAAAT
TGTCTTGGGCCGAAGCTAGAGCATTGATGTTGGCAAGATTGAACTCAATCTTGCAAGGT
GCATCTGGTGCCTCACCTGAAACAATCGACAGAATTGTTGCTGTCTTAAACGCTGGTTT
CGCACCAGAAGTCCCTGCCCAAGGTACTGTAGGTGCTTCCGGTGACTTGACACCATTGG
CACATATGGTTTTGGCCTTACAAGGTAGAGGTAGAATGATTGATCCTAGTGGTAGAGTT
CAAGAAGCCGGTGCTGTCATGGACAGATTATGTGGTGGTCCATTGACTTTAGCTGCAAG
AGATGGTTTGGCTTTAGTTAATGGTACTTCTGCCATGACAGCTATCGCCGCTTTGACAG
GTGTTGAAGCAGCCAGAGCTATTGATGCTGCATTAAGACATTCCGCAGTATTAATGGAA
GTTTTGAGTGGTCATGCAGAAGCCTGGCACCCAGCTTTTGCAGAATTAAGACCACACCC
TGGTCAATTAAGAGCTACCGAAAGATTAGCCCAAGCTTTGGATGGTGCAGGTAGAGTTT
GCAGAACCTTGACTGCCGCTAGAAGATTGACAGCAGCCGACTTAAGACCAGAAGATCAT
CCTGCACAAGACGCCTATTCTTTGAGAGTTGTCCCACAATTAGTTGGTGCTGTCTGGGA
TACTTTGGACTGGCACGATAGAGTAGTTACCTGTGAATTGAACTCAGTCACTGATAACC
CAATATTTCCTGAAGGTTGCGCTGTACCTGCATTACATGGTGGTAATTTCATGGGTGTA
CACGTTGCATTGGCCTCCGACGCTTTAAACGCTGCATTAGTAACATTGGCTGGTTTAGT
TGAAAGACAAATCGCAAGATTGACCGATGAAAAGTTGAATAAGGGTTTGCCAGCATTTT
TGCATGGTGGTCAAGCAGGTTTACAATCAGGTTTCATGGGTGCTCAAGTTACAGCTACC
GCATTGTTAGCAGAAATGAGAGCCAACGCTACCCCTGTCTCTGTACAATCTTTGTCAAC
TAATGGTGCTAACCAAGATGTCGTATCAATGGGTACTATCGCCGCTAGAAGAGCAAGAG
CCCAATTGTTGCCATTGTCTCAAATCCAAGCAATCTTGGCTTTAGCATTGGCCCAAGCT
ATGGACTTGTTAGATGACCCTGAAGGTCAAGCAGGTTGGTCCTTGACAGCCAGAGACTT
AAGAGATAGAATTAGAGCTGTTAGTCCAGGTTTGAGAGCTGATAGACCTTTAGCAGGTC
ATATAGAAGCAGTCGCACAAGGTTTGAGACATCCATCCGCCGCAGCAGACCCTCCAGCC
TAA

At4CL3

ATGATCACTGCAGCTCTGCACGAACCACAGATTCACAAGCCTACCGATACAAGCGTCGT
CAGCGATGATGTATTACCACATTCTCCACCAACGCCACGAATTTTCCGATCAAAATTAC
CCGACATTGACATACCAAACCACCTACCACTACACACTTACTGCTTCGAAAAGCTATCA
TCTGTTAGCGACAAACCATGTCTAATAGTTGGGAGCACCGGGAAGAGCTACACCTACGG
GGAAACACACCTGATATGTCGAAGAGTCGCTAGCGGGCTATACAAACTAGGAATCAGAA
AGGGAGACGTCATAATGATATTACTACAAAACTCAGCGGAGTTCGTTTTCAGCTTCATG
GGAGCTAGCATGATAGGTGCCGTCTCAACCACCGCAAACCCATTCTACACTTCTCAAGA
GTTATATAAGCAGTTAAAGTCTAGCGGTGCGAAGCTAATCATAACTCACTCTCAATACG
TCGATAAGTTAAAGAACCTAGGTGAAAACCTAACGCTGATAACTACCGATGAACCTACA
CCCGAGAATTGTCTGCCTTTCTCGACACTAATAACCGACGACGAAACAAACCCTTTTCA
AGAAACCGTCGATATAGGGGGAGACGATGCGGCGGCGCTACCTTTCTCATCGGGTACAA
CAGGGCTACCTAAGGGTGTTGTTTTAACACACAAAAGCCTAATAACAAGCGTTGCACAA
CAAGTCGATGGTGATAACCCTAATTTATACCTAAAGTCAAACGACGTCATCCTATGCGT
TCTACCTTTGTTCCATATATACTCTCTAAATAGCGTCCTACTAAATTCACTACGTAGCG
GGGCGACGGTTTTACTAATGCATAAATTTGAGATAGGAGCGCTATTAGATTTAATTCAA
AGACATAGAGTAACAATCGCGGCGTTAGTCCCCCCCCTGGTAATAGCTCTGGCTAAGAA
CCCCACGGTTAACTCATATGATCTTTCGAGCGTTAGATTCGTTTTAAGCGGAGCAGCTC
CACTAGGAAAGGAATTACAAGATAGTTTACGTCGACGTCTACCACAAGCGATATTAGGG
CAGGGTTATGGAATGACGGAGGCAGGTCCTGTATTATCAATGAGCTTAGGGTTCGCTAA
GGAACCCATCCCCACAAAGTCAGGATCATGTGGGACTGTAGTCCGTAACGCAGAGTTAA
AGGTAGTTCACTTAGAGACACGTCTATCTTTAGGTTACAACCAACCAGGAGAGATTTGT
ATACGAGGACAACAGATAATGAAGGAGTACTTAAACGATCCTGAAGCGACTTCAGCAAC
AATCGACGAAGAAGGATGGTTACACACAGGTGACATAGGTTATGTTGATGAAGATGATG

AGATTTTCATTGTTGATCGTTTAAAGGAAGTCATAAAATTCAAGGGGTTTCAGGTCCCT
CCTGCTGAGCTGGAGTCCTTACTGATAAATCACCATTCAATTGCGGATGCAGCTGTTGT
TCCCCAAAATGATGAAGTCGCTGGGGAAGTTCCCGTAGCTTTCGTAGTACGTTCAAATG
GTAATGATATAACTGAAGAAGATGTCAAGGAATATGTTGCGAAGCAGGTAGTATTCTAT
AAAAGATTACACAAAGTCTTCTTTGTTGCTAGCATTCCTAAGTCTCCATCGGGAAAGAT
CCTGAGAAAAGACCTAAAGGCTAAATTATGTTGA

GmCHS5 ATGGTTAGTGTTGAAGAAATACGTCAGGCACAACGTGCAGAAGGCCCTGCCACTGTGAT
GGCTATTGGCACAGCCACTCCTCCCAATTGCGTTGATCAGAGTACATATCCTGACTATT
ATTTTCGTATAACAAATTCCGAACACATGACAGAACTAAAGGAAAAGTTTAAGCGTATG
TGTGATAAAAGCATGATTAAAAAACGATACATGTACCTGAATGAAGAAATACTGAAAGA
AAACCCCAGTGTTTGTGCATATATGGCACCTTCGTTAGATGCAAGGCAAGACATGGTTG
TTATGGAAGTGCCAAAATTAGGAAAGGAAGCTGCAACTAAAGCAATAAAAGAATGGGGA
CAACCGAAATCGAAAATTACACATCTTATATTTTGCACAACTAGTGGAGTGGACATGCC
TGGAGCTGATTATCAGCTAACTAAGCTGCTGGGCTTACGTCCCAGTGTGAAACGTTACA
TGATGTACCAACAAGGCTGCTTTGCCGGAGGCACGGTTTTACGTTTAGCCAAGGACCTA
GCTGAAAATAATAAAGGAGCTCGTGTTTTAGTGGTTTGTTCTGAAATCACAGCAGTGAC
ATTTCGTGGCCCAACTGACACACATTTAGATTCCTTAGTTGGACAAGCCTTATTTGGAG
ATGGAGCAGCCGCTGTGATTGTTGGATCAGACCCCCTGCCAGTTGAAAAACCTTTATTT
CAGTTAGTGTGGACTGCCCAGACTATATTACCAGACAGTGAAGGCGCTATTGATGGACA
CTTACGTGAAGTTGGACTAACTTTTCATCTACTAAAAGATGTTCCTGGACTAATAAGTA
AAAACATTGAAAAAGCCTTAGTTGAAGCCTTTCAACCCTTAGGAATCAGTGATTACAAC
TCTATATTTTGGATTGCACACCCTGGAGGACCCGCAATTTTAGACCAAGTTGAAGCTAA
ACTGGGCTTAAAACCTGAAAAGATGGAAGCTACTAGGCATGTTCTATCCGAATATGGAA
ATATGTCAAGTGCATGTGTTCTGTTTATATTAGATCAAATGCGCAAAAAGTCAATCGAA
AACGGATTAGGCACAACAGGCGAAGGCTTAGACTGGGGAGTTCTGTTTGGATTTGGACC
TGGACTAACTGTTGAAACTGTTGTGCTACGTAGTGTGACTGTGTGA

GmCHI1A ATGGCAACGATATCCGCGGTTCAGGTTGAATTTCTTGAATTTCCAGCGGTTGTTACTTC
ACCAGCCAGTGGCAAAACATATTTTCTAGGCGGCGCAGGCGAAAGGGGATTAACGATTG
AAGGCAAATTTATCAAATTTACAGGCATCGGAGTGTACTTAGAAGATAAAGCGGTTCCA
TCACTAGCCGCTAAATGGAAAGGAAAGACTTCAGAAGAACTGGTTCACACACTACACTT
TTACAGGGATATCATTTCAGGCCCGTTTGAAAAACTGATTAGGGGCTCGAAAATTCTTC
CATTAGCTGGCGCTGAATACTCAAAAAAAGTTATGGAAAATTGCGTTGCACACATGAAA
TCTGTTGGCACTTACGGAGATGCTGAAGCCGCAGCCATTGAAAAATTTGCTGAAGCCTT
CAAGAACGTTAACTTTGCACCTGGAGCCTCTGTTTTTTACAGGCAATCACCTGATGGAA
TATTAGGCTTAAGTTTTTCTGAAGATGCAACAATCCCAGAAAAAGAAGCTGCAGTTATC
GAAAACAAGGCTGTGTCAGCGGCGGTGTTAGAAACAATGATTGGAGAACATGCTGTTAG
TCCTGACCTGAAGCGTAGTTTAGCTTCTCGATTACCTGCGGTGTTATCCCACGGCATTA
TCGTGTGA

Table S.4.2: Overview of all plasmids used in this study. The 500u and 500d refer to the 500 bp up – and downstream of the genomic integration place for the knock-out cassettes. At: *Arabidopsis thaliana*, Rc: *Rhodobacter capsulatus*, Gm: *Glycine max*, Ag: *Ashbya gossypii*, Kl: *Kluyveromyces lactis*, Ca: *Candida albicans*, Sp: *Schizosaccharomyces pombe*, Sc: *Saccharomyces cerevisiae*.

Name	Description	Reference
pBN100	pefADR-pAgTEF1-*CaURA3*-tAgTEF1-pefADR, AmpR	[321]
pUG27	loxP- pAgTEF1-*SpHIS5*-tAgTEF1-loxP, AmpR	[319]
pUG73	loxP-pKlLEU2_*KlLEU2*_tKlLEU2-loxP, AmpR	[319]
pSH47	pGAL1-Cre-tCYC1, *URA3*, AmpR, CEN/ARS	[322]
p414	pTEF1-Cas9-tCYC1, *TRP1*, AmpR, CEN/ARS	[19]
p426	pSNR52-gRNA.CAN1.Y-tSUP4, *URA3*, AmpR, 2µ	[19]
pARO3KO	500u-pefADR-pAgTEF1-*CaURA3*-tAgTEF1-pefADR-500d, AmpR	This study
pPDC5KO	500u-loxP-pAgTEF1-*SpHIS5*-tAgTEF1-loxP-500d, AmpR	This study
pPDC6KO	500u-loxP-pKlLEU2_*KlLEU2*_tKlLEU2-loxP-500d, AmpR	This study
pARO10KO	500u-pefADR-pAgTEF1-*CaURA3*-tAgTEF1-pefADR-500d, AmpR	This study
p426*aeBlue*	pSNR52-pBBa_J23110-*aeBlue*, *URA3*, AmpR, 2µ	This study
p426ARO4	pSNR52-gRNA.$ARO4^{G226S}$-tSUP4, *URA3*, AmpR, 2µ	This study
p426ARO7	pSNR52-gRNA.$ARO7^{G141S}$-tSUP4, *URA3*, AmpR, 2µ	This study
$pARO4^{G226S}$	$ARO4^{G226S}$, AmpR	This study
$pARO7^{G141S}$	$ARO7^{G141S}$, AmpR	This study
$pOEACC1^{S659A,S1157A}$	pScTEF1-$ACC1^{S659A,S1157A}$-tScADH1, *URA3*, AmpR, CEN/ARS	This study
pCoumT	pScTEF1-*RcTal1*-tScADH1, *URA3*, AmpR, CEN/ARS	This study
pCoumPT	pTDH3-*AtPAL1*-tENO1, pPGK1-*AtC4H*-tSynth9, pSAC6-*AtATR1*-tGUO1, pTEF1-*RcTal1*-tADH1, *HIS5*, AmpR, CEN/ARS	This study
pNar	pTDH3-*At4CL3*-tGUO1, pPGK1-*GmCHS5*-tSynth17, pTIF6-*GmCHI1A*-tSynth18, *LEU2*, AmpR, CEN/ARS	This study
pURA3	pURA3-*ScURA3*-tURA3, AmpR, CEN/ARS	This study
pHIS5	pTEF1-*SpHIS5*-tTEF1, AmpR, CEN/ARS	This study
pLEU2	pKlLEU2-*KlLEU2*-tTEF1, AmpR, CEN/ARS	This study

Table S.4.3: Listed are gRNA primers, oligonucleotides for the introduction of amino acid modifications in Aro4p, Aro7p and Acc1p and primers for picking up linear donor DNA (DD). The gRNA sequence and the modified amino acid codons are indicated in bold.

Name	Sequence (5' – 3')
ogRNA_ARO4^{G226S}	GCAGTGAAAGATAAATGATC**TGCTCATTCTCACCATTTCA**GTTTTAGAGCTA GAAATAGC
ogRNA_ARO7^{G141S}	GCAGTGAAAGATAAATGATC**GGTGATGATAAGAATAACTT**GTTTTAGAGCTA GAAATAGC
oARO4^{G226S}_fw	CACCATTTCATG**TCC**GTTACTAAGCATGGTGTTG
oARO7^{G141S}_fw	GGTGATGATAAGAACAATTTTT**TCT**TCTGTTGCCACTAG
oARO7^{G141S}_rv	CATTCTATATCTCTAGTGGCAACAGA**AGA**AAAATTGTTCTTATCATCACCATC
oACC1^{S659A}_fw	ATCATACTGAGGCAATTA**GCT**GACGGTGGTCTTTTGATTGCCATAG
oACC1^{S659A}_rv	GCCTATGGCAATCAAAAGACCACCGTC**AGC**TAATTGCCTCAGTATG
oACC1^{S1157A}_fw	CTAAAATGGGAATGAATAGAGCAGTT**GCT**GTTTCAGATTTGTCATATG
oACC1^{S1157A}_rv	CATATGACAAATCTGAAAC**AGC**AACTGCTCTATTCATTCCCATTTTAG
oARO4^{G226S}DD_fw	TTATTACCGGTAAAGACGACAGAG
oARO4^{G226S}DD_rv	GGCAGTGGTAGAGGAAAGAATG
oARO7^{G141S}DD_fw	AAACCAGAAACTGTTTTAAATC
oARO7^{G141S}DD_rv	TACTCTTCCAACCTTCTTAGC

Table S.4.4: Production titers of p-coumaric acid (p-CA) and cinnamic acid (CA) after 72h of growth in synthetic defined medium. The represented values are the mean and standard error of the mean (sem) (n = 3, biological repeats). Strains sCoumPT13 and sCoumT13 are the negative control strains. Strain genotypes are given in Table 6.2. nd: not detected.

Strain	p-CA (mg/l)	p-CA sem (mg/l)	CA (mg/l)	CA sem (mg/l)
sCoumPT01	32,57	1,09	nd	nd
sCoumPT02	5,31	0,34	nd	nd
sCoumPT03	26,48	1,64	nd	nd
sCoumPT04	31,87	0,72	nd	nd
sCoumPT05	33,86	0,87	nd	nd
sCoumPT06	51,48	1,12	nd	nd
sCoumPT07	33,34	0,38	nd	nd
sCoumPT08	34,30	0,70	nd	nd
sCoumPT09	25,16	0,90	nd	nd
sCoumPT10	59,50	4,02	nd	nd
sCoumPT11	13,71	1,02	nd	nd
sCoumPT12	38,93	1,31	nd	nd
sCoumPT13	nd	nd	nd	nd
sCoumT01	21,01	0,39	nd	nd
sCoumT02	2,87	0,27	nd	nd
sCoumT03	13,00	0,30	nd	nd
sCoumT04	21,26	0,59	nd	nd
sCoumT05	19,76	0,20	nd	nd
sCoumT06	31,05	0,51	nd	nd
sCoumT07	21,93	0,65	nd	nd
sCoumT08	20,82	0,44	nd	nd
sCoumT09	16,78	0,25	nd	nd
sCoumT10	6,30	0,16	nd	nd
sCoumT11	3,54	0,06	nd	nd
sCoumT12	2,34	0,19	nd	nd
sCoumT13	nd	nd	nd	nd

Table S.4.5: Production titers of p-coumaric acid (p-CA) and cinnamic acid (CA) after 72h of growth in Feed-In-Time fed-batch medium. The represented values are the mean and standard error of the mean (sem) (n = 3, biological repeats). Strains sCoumPT13 and sCoumT13 are the negative control strains. Strain genotypes are given in Table 6.2. nd: not detected.

Strain	p-CA (mg/l)	p-CA sem (mg/l)	CA (mg/l)	CA sem (mg/l)
sCoumPT01	92,36	2,37	7,36	0,22
sCoumPT02	37,79	1,64	nd	nd
sCoumPT03	123,07	6,03	10,00	0,17
sCoumPT04	88,06	3,47	8,23	0,27
sCoumPT05	91,19	0,15	7,92	0,07
sCoumPT06	100,91	2,29	22,02	0,11
sCoumPT07	67,73	1,35	4,44	0,15
sCoumPT08	82,71	2,27	11,47	0,02
sCoumPT09	43,93	1,28	nd	nd
sCoumPT10	161,91	4,90	13,93	0,64
sCoumPT11	32,38	0,99	nd	nd
sCoumPT12	78,76	1,99	25,48	0,79
sCoumPT13	nd	nd	nd	nd
sCoumT01	39,09	0,42	nd	nd
sCoumT02	19,30	1,16	nd	nd
sCoumT03	36,29	1,05	nd	nd
sCoumT04	41,65	1,36	nd	nd
sCoumT05	37,87	0,66	nd	nd
sCoumT06	82,37	0,84	nd	nd
sCoumT07	39,16	0,56	nd	nd
sCoumT08	76,70	3,29	nd	nd
sCoumT09	21,78	0,16	nd	nd
sCoumT10	12,69	0,35	nd	nd
sCoumT11	10,91	0,35	nd	nd
sCoumT12	nd	nd	nd	nd
sCoumT13	nd	nd	nd	nd

Table S.4.6: Two-way ANOVA analysis to investigate the effect of DAHP synthase and Chorismate mutase on p-coumaric acid production for strains sCoumT09-12 in synthetic defined medium.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	389,063	3	129,688	1343,536	,000
Intercept	628,910	1	628,910	6515,372	,000
DAHPsyn	102,393	1	102,393	1060,767	,000
CHORmut	221,975	1	221,975	2299,610	,000
DAHPsyn * CHORmut	64,696	1	64,696	670,232	,000
Error	,772	8	,097		
Total	1018,745	12			
Corrected Total	389,835	11			

Table S.4.7: Two-way ANOVA analysis to investigate the effect of DAHP synthase and Chorismate mutase on p-coumaric acid production for strains sCoumT09-12 in Feed-In-Time fed-batch medium.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	718,579	3	239,526	1186,694	,000
Intercept	1544,327	1	1544,327	7651,116	,000
DAHPsyn	299,800	1	299,800	1485,311	,000
CHORmut	416,282	1	416,282	2062,400	,000
DAHPsyn * CHORmut	2,497	1	2,497	12,371	,008
Error	1,615	8	,202		
Total	2264,520	12			
Corrected Total	720,193	11			

Table S.4.8: Metabolized titers of p-coumaric acid (p-CA) and production titers of cinnamic acid (CA) and naringenin (Nar) after 72h of growth in Feed-In-Time fed-batch medium. For all strains a final concentration of 164.05 mg/l (1mM) p-coumaric acid was fed to the production strains. The represented values are the mean and standard error of the mean (sem) (n = 3, biological repeats). Strains sNarC04 and sNarAC04 are the negative control strains. Strain genotypes are given in Table 6.2. nd: not detected.

Strain	p-CA (mg/l)	p-CA sem (mg/l)	CA (mg/l)	CA sem (mg/l)	Nar (mg/l)	Nar sem (mg/l)
sNarC01	33,43	0,34	nd	nd	2,00	0,05
sNarC02	24,17	0,38	nd	nd	3,01	0,04
sNarC03	27,29	1,10	nd	nd	5,85	0,29
sNarC04	1,54	0,45	nd	nd	nd	nd
sNarAC01	42,47	0,80	nd	nd	3,25	0,05
sNarAC02	41,71	0,17	nd	nd	5,04	0,17
sNarAC03	39,56	1,46	nd	nd	12,96	0,62
sNarAC04	3,78	0,18	nd	nd	nd	nd

Table S.4.9: Naringenin (Nar) yields obtained after feeding with 164.05 mg/l (1mM) p-coumaric acid (p-CA). The theoretical yield is 1.0 mol Nar/mol p-CA. The represented values are the mean and standard error of the mean (sem) (n = 3, biological repeats). Strain genotypes are given in Table 6.2.

Strain	Yield (mol Nar/mol p-CA)	Yield sem (mol Nar/mol p-CA)
sNarC01	0.036	0.002
sNarC02	0.075	0.001
sNarC03	0.129	0.008
sNarAC01	0.046	0.001
sNarAC02	0.073	0.003
sNarAC03	0.197	0.012

Table S.4.10: *De novo* production titers of p-coumaric acid (p-CA), cinnamic acid (CA), naringenin (Nar) and phloretic acid (Phlor) after 72h of growth in synthetic defined (A) and Feed-In-Time fed-batch medium (B). The represented values are the mean and standard error of the mean (sem) (n = 3, biological repeats). Strains sNarC04 and sNarAC04 are the negative control strains. Strain genotypes are given in Table 6.2. nd: not detected.

A

Strain	p-CA (mg/l)	p-CA sem (mg/l)	CA (mg/l)	CA sem (mg/l)	Nar (mg/l)	Nar sem (mg/l)	Phlor (mg/l)	Phlor sem (mg/l)
sNar01	0,89	0,08	nd	nd	2,33	0,08	16,71	0,12
sNar02	4,76	0,88	nd	nd	4,07	0,24	25,79	0,58
sNar03	8,28	1,84	nd	nd	1,33	0,91	24,69	0,02
sNar04	nd	nd	nd	nd	nd	nd	nd	nd
sNarA01	0,44	0,24	nd	nd	nd	nd	7,62	0,57
sNarA02	3,59	0,28	nd	nd	1,84	0,33	11,08	0,49
sNarA03	14,95	0,24	nd	nd	3,83	0,17	19,30	0,33
sNarA04	nd	nd	nd	nd	nd	nd	nd	nd

B

Strain	p-CA (mg/l)	p-CA sem (mg/l)	CA (mg/l)	CA sem (mg/l)	Nar (mg/l)	Nar sem (mg/l)	Phlor (mg/l)	Phlor sem (mg/l)
sNar01	29,55	0,34	3,08	0,06	0,42	0,06	33,30	0,83
sNar02	62,46	0,79	7,47	0,88	2,00	0,13	37,47	2,23
sNar03	24,97	0,84	4,48	0,13	1,92	0,73	45,66	1,04
sNar04	nd	nd	nd	nd	nd	nd	nd	nd
sNarA01	21,79	0,41	nd	nd	nd	nd	47,11	0,21
sNarA02	42,33	1,71	nd	nd	1,87	0,35	29,90	1,43
sNarA03	37,07	0,50	3,71	0,12	2,92	0,10	29,93	0,65
sNarA04	nd	nd	nd	nd	nd	nd	nd	nd

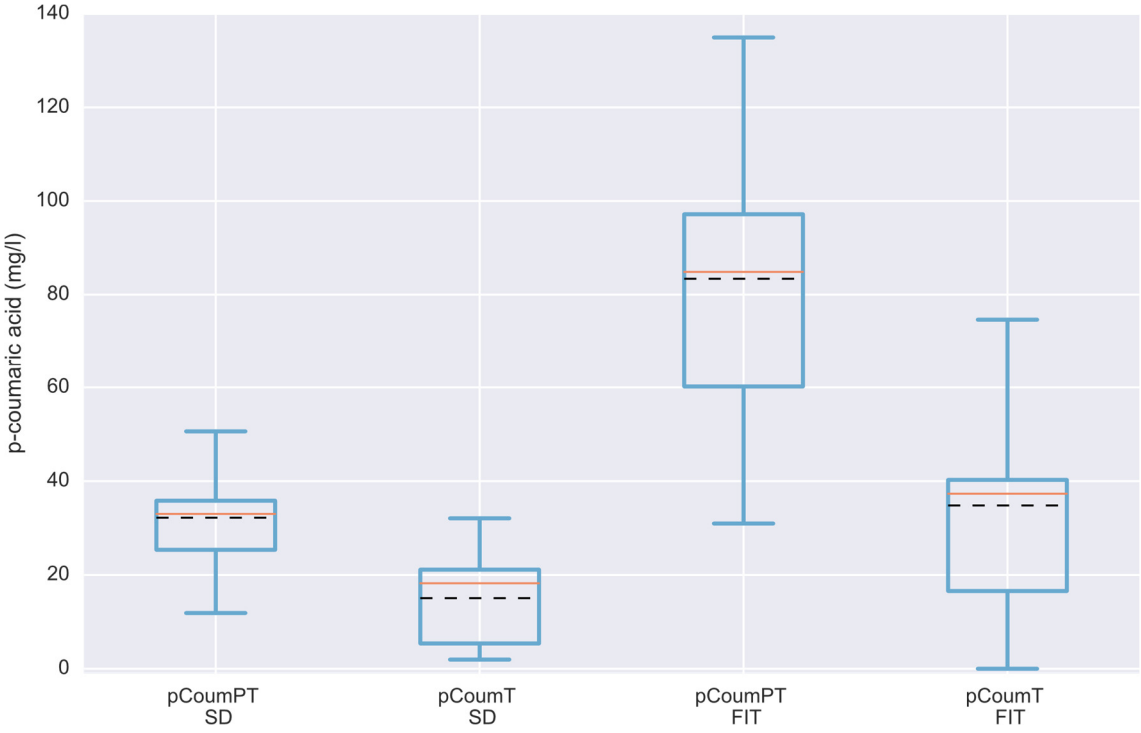

Figure S.4.1: Boxplots of the sCoumPT strains carrying pCoumPT and the sCoumT strains holding pCoumT grown in either synthetic defined (SD) or Feed-In-Time (FIT) fed-batch medium. The full red line represents the median, while the black dotted line indicates the mean. The negative controls sCoumPT13 and sCoumT13 were not taken into account.

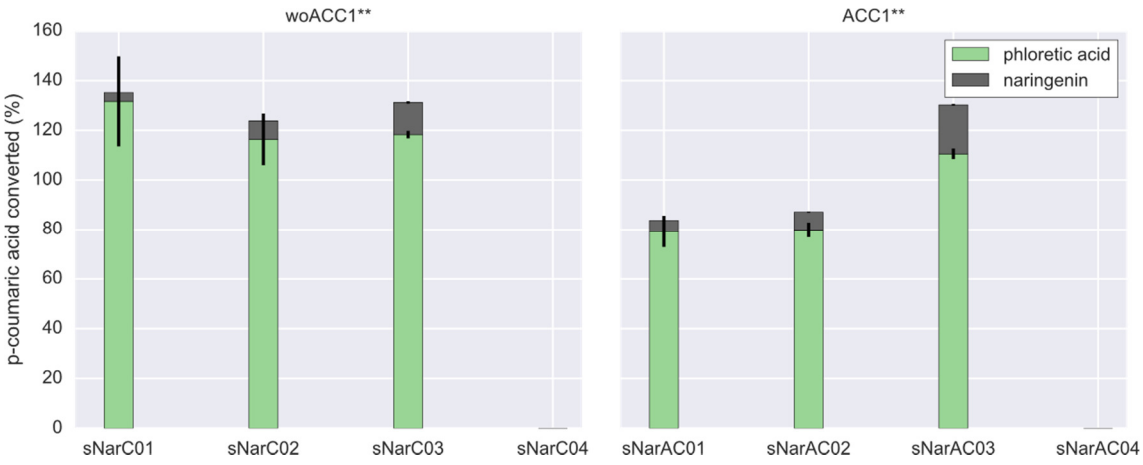

Figure S.4.2: C-balances for strains grown in Feed-In-Time fed-batch medium after 72h when fed with a final concentration of 164.05 mg/l (1mM) p-coumaric acid. Error bars represent the standard error of the mean (n = 3, biological repeats). Strains sNarC01 and sNarAC01 were used as reference strains, sNarC04 and sNarAC04 were the negative control strains. Strain genotypes are given in Table 6.2. Overall, p-coumaric acid is mainly converted to naringenin and phloretic acid. Since non-ideal chromatogram peaks for integration of phloretic acid, quantification is mostly overestimated for this compound. woACC1**: without pOEACC1^{S659A,S1157A}; ACC1**: with pOEACC1^{S659A,S1157A}.

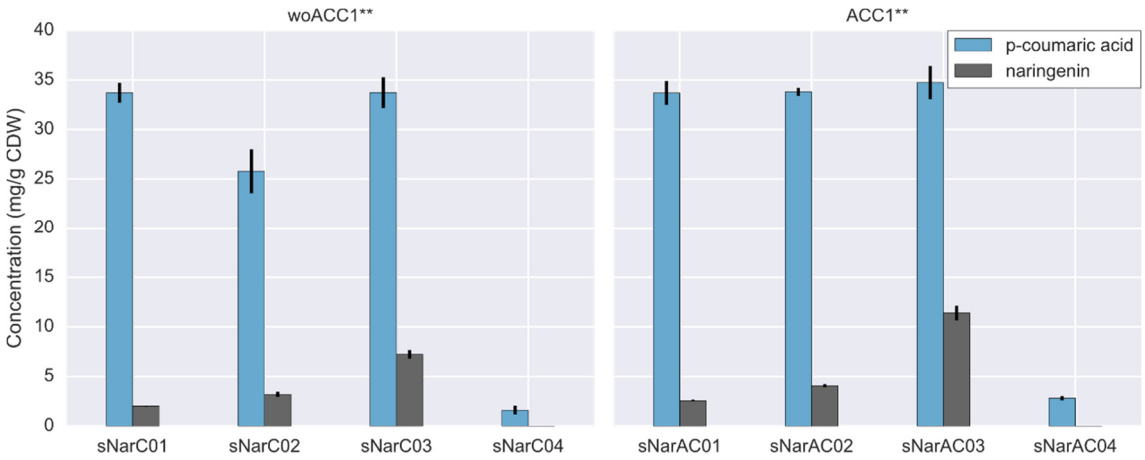

Figure S.4.3: Effect of the overexpression of an improved acetyl-CoA carboxylase (ACC1p^{S659A,S1157A}) on the production titers of naringenin corrected for cell dry weight (CDW) in Feed-In-Time fed-batch medium after 72h. In all strains a final concentration of 164.05 mg/l (1mM) p-coumaric acid was fed to the production strains. The p-coumaric acid concentrations represent the amount that is metabolized, the naringenin titers represent the amount that is produced. Error bars represent the standard error of the mean (n = 3, biological repeats). Strains sNarC01 and sNarAC01 were used as reference strains, sNarC04 and sNarAC04 were the negative control strains. Strain genotypes are given in Table 6.2. woACC1**: without pOEACC1^{S659A,S1157A}; ACC1**: with pOEACC1^{S659A,S1157A}.

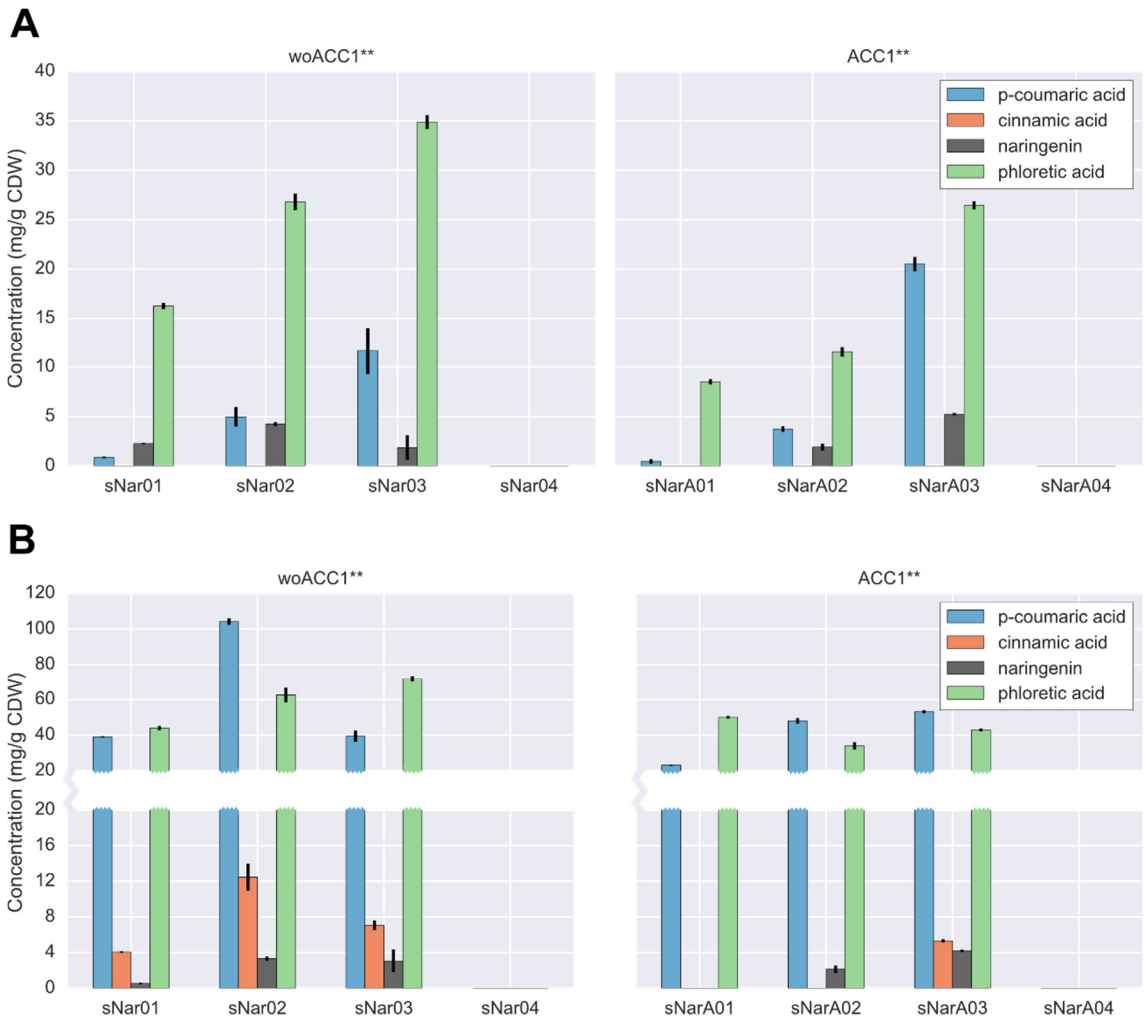

Figure S.4.4: *De novo* production of naringenin corrected for cell dry weight (CDW) in yeast strains with an improved pool of p-coumaric acid whether or not completed with an increased malonyl-CoA pool. Strains were grown for 72h in either synthetic defined (A) or Feed-In-Time fed-batch medium (B). Error bars represent the standard error of the mean (n = 3, biological repeats). Strains sNar01 and sNarA01 were used as reference strains, sNar04 and sNarA04 were the negative control strains. Strain genotypes are given in Table 6.2. woACC1**: without pOEACC1^{S659A,S1157A}; ACC1**: with pOEACC1^{S659A,S1157A}.

Calculation of Cell Dry Weight based on OD600 measurements

The correlation between OD600 measured in a TECAN Infinite® 200 PRO (Tecan) MTP reader and the cell dry weight (CDW) was determined by a dilution range of different optical densities and their corresponding CDW.

$$CDW\ (g/l) = 1{,}4366 \times OD600 - 0{,}0594$$

BIBLIOGRAPHY

1. Guo, Z. & Sherman, F. Signals sufficient for 3'-end formation of yeast mRNA. *Mol. Cell. Biol.* **16,** 2772–2776 (1996).

2. Chae, T. U., Choi, S. Y., Kim, J. W., Ko, Y.-S. & Lee, S. Y. Recent advances in systems metabolic engineering tools and strategies. *Curr. Opin. Biotechnol.* **47,** 67–82 (2017).

3. Lee, S. Y. & Kim, H. U. Systems strategies for developing industrial microbial strains. *Nat. Biotechnol.* **33,** 1061–72 (2015).

4. Ro, D.-K. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440,** 940–943 (2006).

5. Li, M. *et al.* De novo production of resveratrol from glucose or ethanol by engineered Saccharomyces cerevisiae. *Metab. Eng.* **32,** 1–11 (2015).

6. Yim, H. *et al.* Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat Chem Biol* **7,** 445–452 (2011).

7. Galanie, S., Thodey, K., Trenchard, I. J., Interrante, M. F. & Smolke, C. D. Complete biosynthesis of opioids in yeast. *Science (80-.).* **349,** 1095–1100 (2015).

8. Coussement, P., Maertens, J., Beauprez, J., Van Bellegem, W. & De Mey, M. One step DNA assembly for combinatorial metabolic engineering. *Metab. Eng.* **23,** 70–77 (2014).

9. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS One* **3,** e3647 (2008).

10. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–345 (2009).

11. Quan, J. & Tian, J. Circular Polymerase Extension Cloning of Complex Gene Libraries and Pathways. *PLoS One* **4,** e6441 (2009).

12. Sarrion-Perdigones, A. *et al.* GoldenBraid: An iterative cloning system for standardized assembly of reusable genetic modules. *PLoS One* **6,** e21622 (2011).

13. Sarrion-Perdigones, A. *et al.* GoldenBraid 2.0: a comprehensive DNA assembly framework for plant synthetic biology. *Plant Physiol.* **162,** 1618–31 (2013).

14. Goffeau, A. *et al.* Life with 6000 Genes. *Science (80-.).* **274,** 546–567 (1996).

15. Siddiqui, M. S., Thodey, K., Trenchard, I. & Smolke, C. D. Advancing secondary metabolite biosynthesis in yeast with synthetic biology tools. *FEMS Yeast Res.* **12,** 144–170 (2012).

16. Avalos, J. L., Fink, G. R. & Stephanopoulos, G. Compartmentalization of metabolic pathways in yeast mitochondria improves the production of branched-chain alcohols. *Nat. Biotechnol.* **31,** 335–41 (2013).

17. Borodina, I. & Nielsen, J. Advances in metabolic engineering of yeast Saccharomyces cerevisiae for production of chemicals. *Biotechnol. J.* **9,** 609–620 (2014).

18. Li, T. *et al.* TAL nucleases (TALNs): Hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res.* **39,** 359–372 (2011).

19. Dicarlo, J. E. *et al.* Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res.* **41,** 4336–4343 (2013).

20. Jensen, N. B. *et al.* EasyClone: method for iterative chromosomal integration of multiple genes in *Saccharomyces cerevisiae*. *FEMS Yeast Res.* **14,** 238–248 (2014).

21. Mitchell, L. a. *et al.* Versatile genetic assembly system (VEGAS) to assemble pathways for expression in S. cerevisiae. *Nucleic Acids Res.* **43,** 6620–30 (2015).

22. Alper, H., Fischer, C., Nevoigt, E. & Stephanopoulos, G. Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 12678–12683 (2005).

23. Dvir, S. *et al.* Deciphering the rules by which 5 ′ -UTR sequences affect protein expression in yeast. *PNAS* **110,** e2792-801 (2013).

24. Curran, K. a. *et al.* Short, Synthetic Terminators for Improved Heterologous Gene Expression in Yeast. *ACS Synth. Biol.* **4,** 824–32 (2015).

25. Peters, G., Coussement, P., Maertens, J., Lammertyn, J. & De Mey, M. Putting RNA to work: Translating RNA fundamentals into biotechnological engineering practice. *Biotechnol. Adv.* **33,** 1829–1844 (2015).

26. Teo, W. S., Hee, K. S. & Chang, M. W. Bacterial FadR and synthetic promoters function as modular fatty acid sensor- regulators in Saccharomyces cerevisiae. *Eng. Life Sci.* **13,** 456–463 (2013).

27. Jullesson, D., David, F., Pfleger, B. & Nielsen, J. Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. *Biotechnol. Adv.* **33,** 1395–402 (2015).

28. Jensen, M. K. & Keasling, J. D. Recent applications of synthetic biology tools for yeast metabolic engineering. *FEMS Yeast Res.* **15,** 1–10 (2015).

29. Fletcher, E., Krivoruchko, A. & Nielsen, J. Industrial systems biology and its impact on synthetic biology of yeast cell factories. *Biotechnol. Bioeng.* **113,** 1164–1170 (2016).

30. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27,** 946–950 (2009).

31. Marienhagen, J. & Bott, M. Metabolic engineering of microorganisms for the synthesis of plant natural products. *J. Biotechnol.* **163,** 166–178 (2013).

32. Arendt, P., Pollier, J., Callewaert, N. & Goossens, A. Synthetic biology for production of natural and new-to-nature terpenoids in photosynthetic organisms. *Plant J.* **87,** 16–37 (2016).

33. Zhang, Y., Nielsen, J. & Liu, Z. Engineering yeast metabolism for production of terpenoids for use as perfume ingredients, pharmaceuticals and biofuels. *FEMS Yeast Res.* **17,** 1–11 (2017).

34. Paddon, C. J. *et al.* High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* **496,** 528–32 (2013).

35. Moses, T. *et al.* Combinatorial biosynthesis of sapogenins and saponins in Saccharomyces cerevisiae using a C-16α hydroxylase from Bupleurum falcatum.

Proc. Natl. Acad. Sci. U. S. A. **111,** 1634–9 (2014).

36. Albertsen, L. *et al.* Diversion of flux toward sesquiterpene production in saccharomyces cerevisiae by fusion of host and heterologous enzymes. *Appl. Environ. Microbiol.* **77,** 1033–1040 (2011).

37. Dai, Z., Liu, Y., Huang, L. & Zhang, X. Production of miltiradiene by metabolically engineered Saccharomyces cerevisiae. *Biotechnol. Bioeng.* **109,** 2845–2853 (2012).

38. Dai, Z. *et al.* Metabolic engineering of Saccharomyces cerevisiae for production of ginsenosides. *Metab. Eng.* **20,** 146–156 (2013).

39. Hawkins, K. M. & Smolke, C. D. Production of benzylisoquinoline alkaloids in Saccharomyces cerevisiae. *Nat. Chem. Biol.* **4,** 1713–1723 (2008).

40. Brown, S., Clastre, M., Courdavault, V. & O'Connor, S. E. De novo production of the plant-derived alkaloid strictosidine in yeast. *Proc. Natl. Acad. Sci.* **112,** 3205–3210 (2015).

41. Deloache, W. C. *et al.* An enzyme-coupled biosensor enables (S)-reticuline production in yeast from glucose. *Nat. Chem. Biol.* **11,** 465–471 (2015).

42. Trenchard, I. J., Siddiqui, M. S., Thodey, K. & Smolke, C. D. De novo production of the key branch point benzylisoquinoline alkaloid reticuline in yeast. *Metab. Eng.* **31,** 74–83 (2015).

43. Falcone Ferreyra, M. L., Rius, S. P. & Casati, P. Flavonoids: biosynthesis, biological functions, and biotechnological applications. *Front. Plant Sci.* **3,** 222 (2012).

44. Shashank, K. & Abhay, K. Chemistry and Biological Activities of Flavonoids: An Overview. *Sci. World J* **4,** 32–48 (2013).

45. Trantas, E., Koffas, M. a. G., Xu, P. & Ververidis, F. When plants produce not enough or at all: metabolic engineering of flavonoids in microbial hosts. *Front. Plant Sci.* **6,** 1–16 (2015).

46. Pandey, R. P., Parajuli, P., Koffas, M. A. G. & Sohng, J. K. Microbial production of natural and non-natural flavonoids: Pathway engineering, directed evolution and systems/synthetic biology. *Biotechnol. Adv.* **34,** 634–662 (2016).

47. Jiang, H., Wood, K. V & Morgan, J. a. Metabolic Engineering of the Phenylpropanoid Pathway in Saccharomyces cerevisiae Metabolic Engineering of the Phenylpropanoid Pathway in Saccharomyces cerevisiae. *Appl. Environ. Microbiol.* **71,** 2962–2969 (2005).

48. Koopman, F. *et al.* De novo production of the flavonoid naringenin in engineered Saccharomyces cerevisiae. *Microb. Cell Fact.* **11,** 155 (2012).

49. Rodriguez, A. *et al.* Metabolic engineering of yeast for fermentative production of flavonoids. *Bioresour. Technol.* **245,** 1645–1654 (2017).

50. Lyu, X., Ng, K. R., Lee, J. L., Mark, R. & Chen, W. N. Enhancement of Naringenin Biosynthesis from Tyrosine by Metabolic Engineering of Saccharomyces cerevisiae. *J. Agric. Food Chem.* **65,** 6638–6646 (2017).

51. Wang, Y. *et al.* Stepwise increase of resveratrol biosynthesis in yeast Saccharomyces cerevisiae by metabolic engineering. *Metab. Eng.* **13,** 455–463 (2011).

52. Watts, K. T., Lee, P. C. & Schmidt-Dannert, C. Exploring recombinant flavonoid biosynthesis in metabolically engineered Escherichia coli. *ChemBioChem* **5,** 500–507 (2004).

53. Kang, S. Y., Kang, J. Y. & Oh, M. J. Antiviral activities of flavonoids isolated from the bark of Rhus verniciflua stokes against fish pathogenic viruses In Vitro. *J. Microbiol.* **50,** 293–300 (2012).

54. Santos, C. N. S., Koffas, M. & Stephanopoulos, G. Optimization of a heterologous pathway for the production of flavonoids from glucose. *Metab. Eng.* **13,** 392–400 (2011).

55. Delmulle, T., de Maeseneire, S. L. & de Mey, M. Challenges in the microbial production of flavonoids. *Phytochem. Rev.* 1–19 (2017). doi:10.1007/s11101-017-9515-3

56. Kang, S.-Y. *et al.* Artificial biosynthesis of phenylpropanoic acids in a tyrosine overproducing Escherichia coli strain. *Microb. Cell Fact.* **11,** 1–9 (2012).

57. Santos, C. N. S., Xiao, W. & Stephanopoulos, G. Rational, combinatorial, and genomic approaches for engineering L-tyrosine production in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 13538–43 (2012).

58. Lütke-Eversloh, T. & Stephanopoulos, G. L-Tyrosine production by deregulated strains of Escherichia coli. *Appl. Microbiol. Biotechnol.* **75,** 103–110 (2007).

59. Juminaga, D. *et al.* Modular engineering of L-tyrosine production in Escherichia coli. *Appl. Environ. Microbiol.* **78,** 89–98 (2012).

60. Zhou, H., Liao, X., Wang, T., Du, G. & Chen, J. Enhanced l-phenylalanine biosynthesis by co-expression of pheAfbr and aroFwt. *Bioresour. Technol.* **101,** 4151–4156 (2010).

61. Wu, J., Zhou, T., Du, G., Zhou, J. & Chen, J. Modular optimization of heterologous pathways for de Novo synthesis of (2S)-Naringenin in escherichia coli. *PLoS One* **9,** 1–9 (2014).

62. Wu, J., Du, G., Zhou, J. & Chen, J. Metabolic engineering of Escherichia coli for (2S)-pinocembrin production from glucose by a modular metabolic strategy. *Metab. Eng.* **16,** 48–55 (2013).

63. Lee, H., Kim, B. G., Kim, M. & Ahn, J. H. Biosynthesis of two flavones, apigenin and genkwanin, in Escherichia coli. *J. Microbiol. Biotechnol.* **25,** 1442–1448 (2015).

64. Luttik, M. a H. *et al.* Alleviation of feedback inhibition in Saccharomyces cerevisiae aromatic amino acid biosynthesis: Quantification of metabolic impact. *Metab. Eng.* **10,** 141–153 (2008).

65. Rodriguez, A., Kildegaard, K. R., Li, M., Borodina, I. & Nielsen, J. Establishment of a yeast platform strain for production of p-coumaric acid through metabolic engineering of aromatic amino acid biosynthesis. *Metab. Eng.* **31,** 181–188 (2015).

66. Gold, N. *et al.* Metabolic engineering of a tyrosine-overproducing yeast platform

using targeted metabolomics. *Microb. Cell Fact.* **14,** 73 (2015).

67. Zha, W., Rubin-Pitel, S. B., Shao, Z. & Zhao, H. Improving cellular malonyl-CoA level in Escherichia coli via metabolic engineering. *Metab. Eng.* **11,** 192–198 (2009).

68. Fowler, Z. L., Gikandi, W. W. & Koffas, M. A. G. Increased malonyl coenzyme A biosynthesis by tuning the Escherichia coli metabolic network and its application to flavanone production. *Appl. Environ. Microbiol.* **75,** 5831–5839 (2009).

69. Cheng, Z., Jiang, J., Wu, H., Li, Z. & Ye, Q. Enhanced production of 3-hydroxypropionic acid from glucose via malonyl-CoA pathway by engineered Escherichia coli. *Bioresour. Technol.* **200,** 897–904 (2016).

70. Leonard, E., Lim, K.-H., Saw, P.-N. & Koffas, M. A. G. Engineering central metabolic pathways for high-level flavonoid production in Escherichia coli. *Appl. Environ. Microbiol.* **73,** 3877–86 (2007).

71. Xu, P., Ranganathan, S., Fowler, Z. L., Maranas, C. D. & Koffas, M. A. G. Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA. *Metab. Eng.* **13,** 578–587 (2011).

72. van Summeren-Wesenhagen, P. V. & Marienhagen, J. Metabolic engineering of Escherichia coli for the synthesis of the plant polyphenol pinosylvin. *Appl. Environ. Microbiol.* **81,** 840–849 (2015).

73. Cao, W. *et al.* Improved pinocembrin production in Escherichia coli by engineering fatty acid synthesis. *J. Ind. Microbiol. Biotechnol.* **43,** 557–566 (2016).

74. Yang, Y., Lin, Y., Li, L., Linhardt, R. J. & Yan, Y. Regulating malonyl-CoA metabolism via synthetic antisense RNAs for enhanced biosynthesis of natural products. *Metab. Eng.* **29,** 217–226 (2015).

75. Wu, J., Du, G., Chen, J. & Zhou, J. Enhancing flavonoid production by systematically tuning the central metabolic pathways based on a CRISPR interference system in Escherichia coli. *Sci. Rep.* **5,** 1–14 (2015).

76. Wang, Y., Chen, H. & Yu, O. A plant malonyl-CoA synthetase enhances lipid content and polyketide yield in yeast cells. *Appl. Microbiol. Biotechnol.* **98,** 5435–5447 (2014).

77. Wattanachaisaereekul, S., Lantz, A. E., Nielsen, M. L. & Nielsen, J. Production of the polyketide 6-MSA in yeast engineered for increased malonyl-CoA supply. *Metab. Eng.* **10,** 246–254 (2008).

78. Shi, S., Chen, Y. & Siewers, V. Improving Production of Malonyl Coenzyme A-Derived Metabolites. *MBio* **5,** e01130-14 (2014).

79. Ajikumar, P. K. *et al.* Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli. *Science* **330,** 70–74 (2010).

80. Paddon, C. J. & Keasling, J. D. Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat Rev Micro* **12,** 355–367 (2014).

81. Liu, H. & Lu, T. Autonomous production of 1,4-butanediol via a de novo biosynthesis pathway in engineered Escherichia coli. *Metab. Eng.* **29,** 135–141 (2015).

82. Temme, K., Zhao, D. & Voigt, C. a. Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. *Proc. Natl. Acad. Sci.* **109,** 7085–7090 (2012).

83. Smanski, M. J. *et al.* Functional optimization of gene clusters by combinatorial design and assembly. *Nat. Biotechnol.* **32,** 1241–1249 (2014).

84. Baker, M. Is there a reproducibility crisis? *Nature* **533,** 452–454 (2016).

85. Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. The Economics of Reproducibility in Preclinical Research. *PLOS Biol.* **13,** e1002165 (2015).

86. Mutalik, V. K. *et al.* Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat. Methods* **10,** 347–53 (2013).

87. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 14024–14029 (2013).

88. Mutalik, V. K., Qi, L., Guimaraes, J. C., Lucks, J. B. & Arkin, A. P. Rationally designed families of orthogonal RNA regulators of translation. *Nat. Chem. Biol.* **8,** 447–454 (2012).

89. Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10,** 354–60 (2013).

90. Cambray, G. *et al.* Measurement and modeling of intrinsic transcription terminators. *Nucleic Acids Res.* **41,** 5139–5148 (2013).

91. Chen, Y.-J. *et al.* Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat. Methods* **10,** 659–64 (2013).

92. Nielsen, A. A. K. *et al.* Genetic circuit design automation. *Science* **352,** aac7341 (2016).

93. Galdzicki, M. *et al.* Synthetic Biology Open Language (SBOL) Version 1.1.0. 1–26 (2012). doi:1721.1/73909

94. Bartley, B. *et al.* BBF RFC 107: Synthetic Biology Open Language (SBOL) Version 2.0.0. 1–89 (2015).

95. Munson, M., Munro, S. & Salit, M. *Synthetic Biology Standards Consortium Kick off Workshop Report*. (2015).

96. Hayden, E. C. Synthetic biology called to order. *Nature* **520,** 141–142 (2015).

97. Lucks, J. B., Qi, L., Whitaker, W. R. & Arkin, A. P. Toward scalable parts families for predictable design of biological circuits. *Curr. Opin. Microbiol.* **11,** 567–73 (2008).

98. Kelly, J. R. *et al.* Measuring the activity of BioBrick promoters using an in vivo reference standard. *J. Biol. Eng.* **3,** 1–13 (2009).

99. Canton, B., Labno, A. & Endy, D. Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol.* **26,** 787–793 (2008).

100. Hawley, D. K. & Mcclure, W. R. Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Res.* **11,** 2237–2255 (1983).

101. Harley, C. B. & Reynolds, R. P. Analysis of E.coli pormoter sequences. *Nucleic Acids Res.* **15,** 2343–2361 (1987).

102. Lisser, S. & Margalit, H. Compilation of E.coli mRNA promoter sequences. *Nucleic Acids Res.* **21,** 1507–1516 (1993).

103. Weller, K. & Recknagel, R. D. Promoter strength prediction based on occurrence frequencies of consensus patterns. *J. Theor. Biol.* **171,** 355–9 (1994).

104. Goldstein, M. A. & Doi, R. H. Prokaryotic promoters in biotechnology. *Biotechnol. Annu. Rev.* **1,** 105–28 (1995).

105. Ross, W., Aiyar, S. E., Salomon, J. & Gourse, R. L. Escherichia coli promoters with up elements of different strengths: Modular structure of bacterial promoters. *J. Bacteriol.* **180,** 5375–5383 (1998).

106. Burr, T., Mitchell, J., Kolb, A., Minchin, S. & Busby, S. DNA sequence elements located immediately upstream of the -10 hexamer in Escherichia coli promoters: a systematic study. *Nucleic Acids Res.* **28,** 1864–1870 (2000).

107. Mitchell, J. E., Zheng, D., Busby, S. J. W. & Minchin, S. D. Identification and analysis of 'extended -10' promoters in Escherichia coli. *Nucleic Acids Res.* **31,** 4689–4695 (2003).

108. DeMey, M., Maertens, J., Lequeux, G. J., Soetaert, W. K. & Vandamme, E. J. Construction and model-based analysis of a promoter library for E. coli: an indispensable tool for metabolic engineering. *BMC Biotechnol.* **7,** 34 (2007).

109. Brewster, R. C., Jones, D. L. & Phillips, R. Tuning Promoter Strength through RNA Polymerase Binding Site Design in Escherichia coli. *PLoS Comput. Biol.* **8,** e1002811 (2012).

110. Blazeck, J., Garg, R., Reed, B. & Alper, H. S. Controlling promoter strength and regulation in Saccharomyces cerevisiae using synthetic hybrid promoters. *Biotechnol. Bioeng.* **109,** 2884–2895 (2012).

111. Redden, H. & Alper, H. S. The development and characterization of synthetic minimal yeast promoters. *Nat. Commun.* **6,** 7810 (2015).

112. Curran, K. a *et al.* Design of synthetic yeast promoters via tuning of nucleosome architecture. *Nat. Commun.* **5,** 4002 (2014).

113. Lee, M. E., DeLoache, W. C., Cervantes, B. & Dueber, J. E. A Highly-characterized Yeast Toolkit for Modular, Multi-part Assembly. *ACS Synth. Biol.* **4,** 976–986 (2015).

114. Horak, C. E. & Snyder, M. ChIP-chip: A genomic approach for identifying transcription factor binding sites. *Methods Enzymol.* **350,** 469–483 (2002).

115. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10,** 669–80 (2009).

116. Vilanova, C. *et al.* Standards not that standard. *J. Biol. Eng.* **9,** 17 (2015).

117. VanHove, B., Love, A. M., Ajikumar, P. K. & De Mey, M. in *Synthetic Biology* 1–64 (2016). doi:10.1007/978-3-319-22708-5

118. Englaender, J. A. *et al.* Effect of Genomic Integration Location on Heterologous Protein Expression and Metabolic Engineering in E. coli. *ACS Synth. Biol.* **6,** 710–720

(2017).

119. Sauer, C. *et al.* Effect of Genome Position on Heterologous Gene Expression in Bacillus subtilis: An Unbiased Analysis. *ACS Synth. Biol.* **5,** 942–947 (2016).

120. Flagfeldt, D. B., Siewers, V., Huang, L. & Nielsen, J. Characterization of chromosomal integration sites for heterologous gene expression in Saccharomyces cerevisiae. *Yeast* **26,** 545–551 (2009).

121. Gilman, J. & Love, J. Synthetic promoter design for new microbial chassis. *Biochem. Soc. Trans.* **44,** 731–737 (2016).

122. Davis, J. H., Rubin, A. J. & Sauer, R. T. Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic Acids Res.* **39,** 1131–1141 (2011).

123. Lou, C., Stanton, B., Chen, Y.-J., Munsky, B. & Voigt, C. A. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat. Biotechnol.* **30,** 1137–42 (2012).

124. Carr, S. B., Beal, J. & Densmore, D. M. Reducing DNA context dependence in bacterial promoters. *PLoS One* **12,** e0176013 (2017).

125. Casini, A., Storch, M., Baldwin, G. S. & Ellis, T. Bricks and blueprints: methods and standards for DNA assembly. *Nat. Rev. Mol. Cell Biol.* **16,** 568–76 (2015).

126. Martínez-Garćía, E., Aparicio, T., Goñi-Moreno, A., Fraile, S. & De Lorenzo, V. SEVA 2.0: An update of the Standard European Vector Architecture for de-/re-construction of bacterial functionalities. *Nucleic Acids Res.* **43,** D1183–D1189 (2015).

127. Knight, T. Idempotent Vector Design for Standard Assembly of Biobricks. *MIT Synth. Biol. Work. Gr. Tech. Reports* 1–11 (2003). doi:http://hdl.handle.net/1721.1/21168

128. Shetty, R. P., Endy, D. & Knight, T. F. Engineering BioBrick vectors from BioBrick parts. *J. Biol. Eng.* **2,** 1–12 (2008).

129. Anderson, J. C. *et al.* BglBricks: A flexible standard for biological part assembly. *J. Biol. Eng.* **4,** 1–12 (2010).

130. Sarrion-Perdigones, A. *et al.* GoldenBraid: an iterative cloning system for standardized assembly of reusable genetic modules. *PLoS One* **6,** e21622 (2011).

131. Van Hove, B., Guidi, C., De Wannemaeker, L., Maertens, J. & De Mey, M. Recursive DNA Assembly Using Protected Oligonucleotide Duplex Assisted Cloning (PODAC). *ACS Synth. Biol.* **6,** 943–949 (2017).

132. Ellis, T., Adie, T. & Baldwin, G. S. DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr. Biol. (Camb).* **3,** 109–118 (2011).

133. Jin, P., Ding, W., Du, G., Chen, J. & Kang, Z. DATEL: A Scarless and Sequence-Independent DNA Assembly Method Using Thermostable Exonucleases and Ligase. *ACS Synth. Biol.* **5,** 1028–1032 (2016).

134. Liang, J., Liu, Z., Low, X. Z., Ang, E. L. & Zhao, H. Twin-primer non-enzymatic DNA assembly: an efficient and accurate multi-part DNA assembly method. *Nucleic Acids Res.* **45,** e94 (2017).

135. Zhou, J., Wu, R., Xue, X. & Qin, Z. CasHRA (Cas9-facilitated Homologous Recombination Assembly) method of constructing megabase-sized DNA. *Nucleic Acids Res.* **44,** e124 (2016).

136. Beal, J. *et al.* Reproducibility of Fluorescent Expression from Engineered Biological Constructs in E. coli. *PLoS One* **11,** 1–22 (2016).

137. Linshiz, G. *et al.* PaR-PaR laboratory automation platform. *ACS Synth. Biol.* **2,** 216–222 (2013).

138. Shih, S. C. C. *et al.* A Versatile Microfluidic Device for Automating Synthetic Biology. *ACS Synth. Biol.* **4,** 1151–1164 (2015).

139. Linshiz, G. *et al.* End-to-end automated microfluidic platform for synthetic biology: from design to functional analysis. *J. Biol. Eng.* **10,** 1–15 (2016).

140. Goni-Moreno, A. *et al.* An Implementation-Focused Bio/Algorithmic Workflow for Synthetic Biology. *ACS Synth. Biol.* **5,** 1127–1135 (2016).

141. Myers, C. J. *et al.* A standard-enabled workflow for synthetic biology. *Biochem. Soc. Trans.* **45,** 793–803 (2017).

142. Zhang, M., McLaughlin, J. A., Wipat, A. & Myers, C. J. SBOLDesigner 2: An Intuitive Tool for Structural Genetic Design. *ACS Synth. Biol.* **6,** 1150–1160 (2017).

143. Shaner, N. C., Steinbach, P. a & Tsien, R. Y. A guide to choosing fluorescent proteins. *Nat. Methods* **2,** 905–909 (2005).

144. Delvigne, F., Pêcheux, H. & Tarayre, C. Fluorescent Reporter Libraries as Useful Tools for Optimizing Microbial Cell Factories: A Review of the Current Methods and Applications. *Front. Bioeng. Biotechnol.* **3,** 1–8 (2015).

145. Hebisch, E., Knebel, J., Landsberg, J., Frey, E. & Leisner, M. High Variation of Fluorescence Protein Maturation Times in Closely Related Escherichia coli Strains. *PLoS One* **8,** e75991 (2013).

146. Shin, I. *et al.* Live-cell imaging of Pol II promoter activity to monitor gene expression with RNA IMAGEtag reporters. *Nucleic Acids Res.* **42,** e90 (2014).

147. Pothoulakis, G., Ceroni, F., Reeve, B. & Ellis, T. The Spinach RNA aptamer as a characterisation tool for synthetic biology. *ACS Synth. Biol.* **3,** 182–187 (2013).

148. Filonov, G. S., Moon, J. D. & Svensen, N. Broccoli: Rapid Selection of an RNA Mimic of Green Fluorescent Protein by Fluorescence-Based Selection and Directed Evolution. *J. Am. Chem. Soc.* **136,** 16299–16308 (2014).

149. Chappell, J., Jensen, K. & Freemont, P. S. Validation of an entirely in vitro approach for rapid prototyping of DNA regulatory elements for synthetic biology. *Nucleic Acids Res.* **41,** 3471–3481 (2013).

150. Smith, M. T., Wilding, K. M., Hunt, J. M., Bennett, A. M. & Bundy, B. C. The emerging age of cell-free synthetic biology. *FEBS Lett.* **588,** 2755–2761 (2014).

151. Siegal-Gaskins, D., Tuza, Z. A., Kim, J., Noireaux, V. & Murray, R. M. Resource usage and gene circuit performance characterization in a cell-free 'breadboard'. *ACS Synth. Biol.*

3, 416–425 (2014).

152. Garamella, J., Marshall, R., Rustad, M. & Noireaux, V. The All E. coli TX-TL Toolbox 2.0: A Platform for Cell-Free Synthetic Biology. *ACS Synth. Biol.* **5,** 344–355 (2016).

153. Moore, S. J. *et al.* EcoFlex: A Multifunctional MoClo Kit for E. coli Synthetic Biology. *ACS Synth. Biol.* **5,** 1059–1069 (2016).

154. Li, J. *et al.* Green fluorescent protein in Saccharomyces cerevisiae: Real-time studies of the GAL 1 promoter. *Biotechnol. Bioeng.* **70,** 187–196 (2000).

155. Huber, R., Roth, S., Rahmen, N. & Büchs, J. Utilizing high-throughput experimentation to enhance specific productivity of an E.coli T7 expression system by phosphate limitation. *BMC Biotechnol.* **11,** 22 (2011).

156. Song, Y. *et al.* Promoter screening from Bacillus subtilis in various conditions hunting for synthetic biology and industrial applications. *PLoS One* **11,** e0158447 (2016).

157. Blazeck, J., Liu, L., Redden, H. & Alper, H. Tuning gene expression in yarrowia lipolytica by a hybrid promoter approach. *Appl. Environ. Microbiol.* **77,** 7905–7914 (2011).

158. Mishra, S., Anand, D., Vijayarangan, N. & Ajitkumar, P. An accurate method for the qualitative detection and quantification of mycobacterial promoter activity. *Open Microbiol. J.* **7,** 1–5 (2013).

159. Mumberg, D., Müller, R. & Funk, M. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene* **156,** 119–122 (1995).

160. Zwietering, M. H., Jongenburger, I., Rombouts, F. M. & van 't Riet, K. Modeling of the Bacterial Growth Curve. *Appl. Environ. Microbiol.* **56,** 1875–1881 (1990).

161. Miller, J. & Fenton, S. Shares of sweetener maker Evolva sour after revenue warning. Available at: http://af.reuters.com/article/commodities07News/idAFL5N1F01JW. (Accessed: 1st August 2017)

162. Segall-Shapiro, T. H., Meyer, A. J., Ellington, A. D., Sontag, E. D. & Voigt, C. a. A 'resource allocator' for transcription based on a highly fragmented T7 RNA polymerase. *Mol. Syst. Biol.* **10,** 742 (2014).

163. Mashego, M. R., Jansen, M. L. A., Vinke, J. L., Van Gulik, W. M. & Heijnen, J. J. Changes in the metabolome of Saccharomyces cerevisiae associated with evolution in aerobic glucose-limited chemostats. *FEMS Yeast Res.* **5,** 419–430 (2005).

164. Shis, D. L. & Bennett, M. R. Library of synthetic transcriptional AND gates built with split T7 RNA polymerase mutants. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 5028–33 (2013).

165. Lee, Y., Lafontaine Rivera, J. G. & Liao, J. C. Ensemble Modeling for Robustness Analysis in engineering non-native metabolic pathways. *Metab. Eng.* **25,** 63–71 (2014).

166. Field, D. *et al.* The minimum information about a genome sequences (MIGS) specification. *Nat Biotechnol.* **26,** 541–547 (2008).

167. Bustin, S. A. *et al.* The MIQE Guidelines: Minimum Information for Publication of

Quantitative Real-Time PCR Experiments. *Clin. Chem.* **55,** 611–622 (2009).

168. Jenkins, H. *et al.* A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.* **22,** 1601–1606 (2004).

169. Dondrup, M. *et al.* EMMA 2 - a MAGE-compliant system for the collaborative analysis and integration of microarray data. *BMC Bioinformatics* **10,** 1–14 (2009).

170. Robin, X., Hoogland, C., Appel, R. D. & Lisacek, F. MIAPEGelDB, a web-based submission tool and public repository for MIAPE gel electrophoresis documents. *J. Proteomics* **71,** 249–251 (2008).

171. Chavez, M., Ho, J. & Tan, C. Reproducibility of high-throughput plate-reader experiments in synthetic biology. *ACS Synth. Biol.* **6,** 375–380 (2016).

172. Misirli, G. *et al.* Data Integration and Mining for Synthetic Biology Design. *ACS Synth. Biol.* **5,** 1086–1097 (2016).

173. Sainz de Murieta, I., Bultelle, M. & Kitney, R. I. Towards the first data acquisition standard in Synthetic Biology. *ACS Synth. Biol.* **5,** 817–826 (2016).

174. Huynh, L. & Tagkopoulos, I. A Parts Database with Consensus Parameter Estimation for Synthetic Circuit Design. *ACS Synth. Biol.* **5,** 1412–1420 (2016).

175. Hillson, N. J., Plahar, H. A., Beal, J. & Prithviraj, R. Improving Synthetic Biology Communication: Recommended Practices for Visual Depiction and Digital Submission of Genetic Designs. *ACS Synth. Biol.* **5,** 449–51 (2016).

176. Peccoud, J. *et al.* Essential information for synthetic DNA sequences. *Nat. Publ. Gr.* **29,** 22 (2011).

177. Polat, K. & Güneş, S. A novel approach to estimation of E. coli promoter gene sequences: Combining feature selection and least square support vector machine (FS_LSSVM). *Appl. Math. Comput.* **190,** 1574–1582 (2007).

178. Meng, H., Ma, Y., Mai, G., Wang, Y. & Liu, C. Construction of precise support vector machine based models for predicting promoter strength. *Quant. Biol.* **5,** 90–98 (2017).

179. Na, D. & Lee, D. RBSDesigner: Software for designing synthetic ribosome binding sites that yields a desired level of protein expression. *Bioinformatics* **26,** 2633–2634 (2010).

180. Seo, S. W. *et al.* Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metab. Eng.* **15,** 67–74 (2013).

181. Bonde, M. T. *et al.* Predictable tuning of protein expression in bacteria. *Nat. Methods* **13,** 233–236 (2016).

182. Espah Borujeni, A., Mishler, D. M., Wang, J., Huso, W. & Salis, H. M. Automated physics-based design of synthetic riboswitches from diverse RNA aptamers. *Nucleic Acids Res.* **44,** 1–13 (2015).

183. Roehner, N., Young, E. M., Voigt, C. A., Gordon, D. B. & Densmore, D. Double Dutch: A Tool for Designing Combinatorial Libraries of Biological Systems. *ACS Synth. Biol.* **5,**

507–17 (2016).

184. Rodrigo, G. & Jaramillo, A. AutoBioCAD: Full biodesign automation of genetic circuits. *ACS Synth. Biol.* **2,** 230–236 (2013).

185. Otero-Muras, I. & Banga, J. R. Automated Design Framework for Synthetic Biology exploiting Pareto Optimality. *ACS Synth. Biol.* **6,** 1180–1193 (2017).

186. Davidsohn, N., Beal, J., Adler, A., Yaman, F. & Li, Y. Accurate predictions of genetic circuit behavior from part characterization and modular composition. *ACS Synth. Biol.* **4,** 673–681 (2013).

187. Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. *Cell* **164,** 1185–1197 (2016).

188. Curran, K. a., Karim, A. S., Gupta, A. & Alper, H. S. Use of expression-enhancing terminators in Saccharomyces cerevisiae to increase mRNA half-life and improve gene expression control for metabolic engineering applications. *Metab. Eng.* **19,** 88–97 (2013).

189. MacPherson, M. & Saka, Y. Short Synthetic Terminators for Assembly of Transcription Units in Vitro and Stable Chromosomal Integration in Yeast S. cerevisiae. *ACS Synth. Biol.* **6,** 130–138 (2016).

190. Redden, H., Morse, N. & Alper, H. S. The synthetic biology toolbox for tuning gene expression in yeast. *FEMS Yeast Res.* **15,** 1–10 (2015).

191. Sun, J. *et al.* Cloning and characterization of a panel of constitutive promoters for applications in pathway engineering in Saccharomyces cerevisiae. *Biotechnol. Bioeng.* **109,** 2082–2092 (2012).

192. Lee, M. E., DeLoache, W. C., Cervantes, B. & Dueber, J. E. A Highly-characterized Yeast Toolkit for Modular, Multi-part Assembly. *ACS Synth. Biol.* **4,** 975–986 (2015).

193. Da Silva, N. a. & Srikrishnan, S. Introduction and expression of genes for metabolic engineering applications in Saccharomyces cerevisiae. *FEMS Yeast Res.* **12,** 197–214 (2012).

194. Blazeck, J. & Alper, H. S. Promoter engineering: Recent advances in controlling transcription at the most fundamental level. *Biotechnol. J.* **8,** 46–58 (2013).

195. Lubliner, S., Keren, L. & Segal, E. Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Res.* **41,** 5569–5581 (2013).

196. Lubliner, S. *et al.* Core promoter sequence in yeast is a major determinant of expression level. *Genome Res.* **25,** 1008–1017 (2015).

197. Danino, Y. M., Even, D., Ideses, D. & Juven-Gershon, T. The core promoter: at the heart of gene expression. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1847,** 1116–31 (2015).

198. Ede, C., Chen, X., Lin, M.-Y. & Chen, Y. Y. Quantitative Analyses of Core Promoters Enable Precise Engineering of Regulated Gene Expression in Mammalian Cells. *ACS Synth. Biol.* **5,** 395–404 (2016).

199. Portela, R. M. C. *et al.* Synthetic core promoters as universal parts for fine-tuning expression in different yeast species. *ACS Synth. Biol.* **6,** 471–484 (2017).

200. Lin, Z., Wu, W.-S., Liang, H., Woo, Y. & Li, W.-H. The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcriptional regulation. *BMC Genomics* **11,** 581 (2010).

201. Hahn, S. & Young, E. T. Transcriptional regulation in saccharomyces cerevisiae: Transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* **189,** 705–736 (2011).

202. Nevoigt, E. *et al.* Engineering of promoter replacement cassettes for fine-tuning of gene expression in Saccharomyces cerevisiae. *Appl. Environ. Microbiol.* **72,** 5266–5273 (2006).

203. Dower, K. & Rosbash, M. T7 RNA polymerase-directed transcripts are processed in yeast and link 3′ end formation to mRNA nuclear export. *Rna* **8,** 686–697 (2002).

204. Benton, B. M. *et al.* Signal-mediated import of bacteriophage T7 RNA polymerase into the Saccharomyces cerevisiae nucleus and specific transcription of target genes. *Mol. Cell. Biol.* **10,** 353–360 (1990).

205. Brachmann, C. B. *et al.* Designer deletion strains derived from Saccharomyces cerevisiae S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14,** 115–132 (1998).

206. Sheff, M. a. & Thorn, K. S. Optimized cassettes for fluorescent protein tagging in Saccharomyces cerevisiae. *Yeast* **21,** 661–670 (2004).

207. Gay, P., Le Coq, D., Steinmetz, M., Ferrari, E. & Hoch, J. A. Cloning structural gene sacB, which codes for exoenzyme levansucrase of Bacillus subtilis: Expression of the gene in Escherichia coli. *J. Bacteriol.* **153,** 1424–1431 (1983).

208. Gietz, R. D. & Schiestl, R. H. Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2,** 35–37 (2007).

209. Weenink, T., Mckiernan, R. M. & Ellis, T. Rational Design of RNA Structures that Predictably Tune Eukaryotic Gene Expression. *BioRxiv* doi: http://dx.doi.org/10.1101/137877 (2017).

210. Li, S., Si, T., Wang, M. & Zhao, H. Development of a Synthetic Malonyl-CoA Sensor in Saccharomyces cerevisiae for Intracellular Metabolite Monitoring and Genetic Screening. *ACS Synth. Biol.* **4,** 1308–1315 (2015).

211. David, F., Nielsen, J. & Siewers, V. Flux Control at the Malonyl-CoA Node through Hierarchical Dynamic Pathway Regulation in Saccharomyces cerevisiae. *ACS Synth. Biol.* **5,** 224–233 (2016).

212. Jendresen, C. B. *et al.* Highly Active and Specific Tyrosine Ammonia-Lyases from Diverse Origins Enable Enhanced Production of Aromatic Compounds in Bacteria and Saccharomyces cerevisiae. *Appl. Environ. Microbiol.* **81,** 4458–4476 (2015).

213. Cramer, P., Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural Basis of Transcription : RNA Polymerase II at 2 . 8 Ångstrom Resolution. *Science (80-.).* **1863,** 1863–1877 (2001).

214. Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* **19,** 65–71 (2009).

215. Raveh-Sadka, T. *et al.* Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.* **44,** 743–750 (2012).

216. Zeevi, D. *et al.* Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. *Genome Res.* **21,** 2114–2128 (2011).

217. Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.* **15,** 453–68 (2014).

218. Jansen, A., Van Der Zande, E., Meert, W., Fink, G. R. & Verstrepen, K. J. Distal chromatin structure influences local nucleosome positions and gene expression. *Nucleic Acids Res.* **40,** 3870–3885 (2012).

219. Zhang, Z. & Dietrich, F. S. Mapping of transcription start sites in Saccharomyces cerevisiae using 5′ SAGE. *Nucleic Acids Res.* **33,** 2838–2851 (2005).

220. Chen, W. & Struhl, K. Yeast mRNA initiation sites are determined primarily by specific sequences, not by the distance from the TATA element. *EMBO J.* **4,** 3273–80 (1985).

221. Hahn, S., Hoar, E. T. & Guarente, L. Each of three 'TATA elements' specifies a subset of the transcription initiation sites at the CYC-1 promoter of Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **82,** 8562–8566 (1985).

222. Grigaite, R., Maneliene, Z. & Janulaitis, A. AarI, a restriction endonuclease from Arthrobacter aurescens SS2-322, which recognizes the novel non-palindromic sequence 5′-CACCTGC(N)4/8-3′. *Nucleic Acids Res* **30,** e123 (2002).

223. Li, X. Y. *et al.* Selective recruitment of TAFs by yeast upstream activating sequences: Implications for eukaryotic promoter structure. *Curr. Biol.* **12,** 1240–1244 (2002).

224. Hartmann, M. *et al.* Evolution of feedback-inhibited beta /alpha barrel isoenzymes by gene duplication and a single mutation. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 862–867 (2003).

225. Salis, H. M. The ribosome binding site calculator. *Methods Enzymol.* **498,** 19–42 (2011).

226. Hinnebusch, A. G. Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiol. Mol. Biol. Rev.* **75,** 434–467 (2011).

227. Hinnebusch, A. G., Dever, T. E. & Sonenberg, N. Mechanism and Regulation of Protein Synthesis Initiation in Eukaryotes. *Genetics* **203,** 65–107 (2016).

228. Araujo, P. R. *et al.* Before it gets started: Regulating translation at the 5 UTR. *Comp. Funct. Genomics* **2012,** 8 pages (2012).

229. Tuller, T., Ruppin, E. & Kupiec, M. Properties of untranslated regions of the S. cerevisiae genome. *BMC Genomics* **10,** doi:10.1186/1471-2164-10-391 (2009).

230. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).

231. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 5320–5325 (2006).

232. Lin, Z. & Li, W. H. Evolution of 5' untranslated region length and gene expression reprogramming in yeasts. *Mol. Biol. Evol.* **29,** 81–89 (2012).

233. Kozak, M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361,** 13–37 (2005).

234. Robbins-Pianka, A., Rice, M. D. & Weir, M. P. The mRNA landscape at yeast translation initiation sites. *Bioinformatics* **26,** 2651–2655 (2010).

235. Zur, H. & Tuller, T. New Universal Rules of Eukaryotic Translation Initiation Fidelity. *PLoS Comput. Biol.* **9,** e1003136 (2013).

236. Ben-Yehezkel, T., Zur, H., Marx, T., Shapiro, E. & Tuller, T. Mapping the translation initiation landscape of an S. cerevisiae gene using fluorescent proteins. *Genomics* **102,** 419–429 (2013).

237. Ringnér, M. & Krogh, M. Folding free energies of 5'-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput. Biol.* **1,** 585–592 (2005).

238. Zhang, Z. & Dietrich, F. S. Identification and characterization of upstream open reading frames (uORF) in the 5′ untranslated regions (UTR) of genes in Saccharomyces cerevisiae. *Curr. Genet.* **48,** 77–87 (2005).

239. Cuperus, J. *et al.* Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Res.* **27,** 2015–2024 (2017).

240. Ai, H., Henderson, J. N., Remington, S. J. & Campbell, R. E. Directed evolution of a monomeric, bright and photostable version of Clavularia cyan fluorescent protein: structural characterization and applications in fluorescence imaging. *Biochem. J.* **400,** 531–40 (2006).

241. Kyndt, J. A., Meyer, T. E., Cusanovich, M. A. & Beeumen, J. J. Van. Characterization of a bacterial tyrosine ammonia lyase, a biosynthetic enzyme for the photoactive yellow protein. *Febs Lett* **512,** 240–244 (2002).

242. Stansfield, I. & Stark, M. J. R. *Yeast Gene Analysis*. (Elsevier Ltd, 2007).

243. Hofacker, I. L. *et al.* Fast Folding and Comparison of RNA Secondary Structure. *Monatshefte für Chemie* **125,** 167–188 (1994).

244. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6,** 26 (2011).

245. Mevik, B.-H. & Wehrens, R. The pls Package: Principle Component and Partial Least Squares Regression in R. *J. Stat. Softw.* **18,** 1–24 (2007).

246. Dayal, B. S. & MacGregor, J. F. Improved PLS Algorithms. *J. Chemom.* **11,** 73–85 (1997).

247. De Mey, M., Maertens, J., Lequeux, G. J., Soetaert, W. K. & Vandamme, E. J. Construction and model-based analysis of a promoter library for E. coli: an indispensable tool for metabolic engineering. *BMC Biotechnol.* **7,** 34 (2007).

248. Gordan, R. *et al.* Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Rep.* **3,**

1093–1104 (2013).

249. Lee, M. E., Aswani, A., Han, A. S., Tomlin, C. J. & Dueber, J. E. Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic Acids Res.* **41,** 10668–10678 (2013).

250. Rajkumar, A. S., Dénervaud, N. & Maerkl, S. J. Mapping the fine structure of a eukaryotic promoter input-output function. *Nat. Genet.* **45,** 1207–15 (2013).

251. Selpi *et al.* Predicting functional upstream open reading frames in Saccharomyces cerevisiae. *BMC Bioinformatics* **10,** 451 (2009).

252. Miyasaka, H. The positive relationship between codon usage bias and translation initiation AUG context in Saccharomyces cerevisiae. *Yeast* **15,** 633–637 (1999).

253. Yun, Y., Adesanya, T. M. A. & Mitra, R. D. A systematic study of gene expression variation at single nucleotide resolution reveals widespread regulatory roles for uAUGs. *Genome Res.* **22,** 1089–1097 (2012).

254. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (2010).

255. Saglioccos, F. a *et al.* The Influence of 5 ' -Secondary Structures upon Ribosome Binding to mRNA during Translation in Yeast. *J. Biol. Chem.* **268,** 26522–26530 (1993).

256. Lamping, E., Niimi, M. & Cannon, R. D. Small, synthetic, GC-rich mRNA stem-loop modules 5' proximal to the AUG start-codon predictably tune gene expression in yeast. *Microb. Cell Fact.* **12,** 74 (2013).

257. Crook, N. C., Freeman, E. S. & Alper, H. S. Re-engineering multicloning sites for function and convenience. *Nucleic Acids Res.* **39,** e92 (2011).

258. Wang, X. Q. & Rothnagel, J. A. 5-Untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. *Nucleic Acids Res.* **32,** 1382–1391 (2004).

259. Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44,** 283–292 (1986).

260. Nakagawa, S., Niimura, Y., Gojobori, T., Tanaka, H. & Miura, K. ichiro. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res.* **36,** 861–871 (2008).

261. Dacheux, E. *et al.* Translation initiation events on structured eukaryotic mRNAs generate gene expression noise. *Nucleic Acids Res.* **45,** 6981–6992 (2017).

262. Zid, B. M. & O'Shea, E. K. Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast. *Nature* **514,** 117–121 (2014).

263. Morris, D R;Geballe, A. P. Upstream open reading frames as regulators of mRNA translation. *Mol. Cell. Biol.* **20,** 8635–8642 (2000).

264. Meijer, H. A. & Thomas, A. A. M. Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA. *Biochem. J.* **367,** 1–11

(2002).

265. Iacono, M., Mignone, F. & Pesole, G. uAUG and uORFs in human and rodent 5'untranslated mRNAs. *Gene* **349,** 97–105 (2005).

266. Cavener, D. R. & Ray, S. C. Eukaryotic start and stop translation sites. *Nucleic Acids Res.* **19,** 3185–3192 (1991).

267. Li, M., Schneider, K., Kristensen, M., Borodina, I. & Nielsen, J. Engineering yeast for high-level production of stilbenoid antioxidants. *Sci. Rep.* **6,** 36827 (2016).

268. Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S. & Vale, R. D. A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell* **159,** 635–646 (2014).

269. Yan, X., Hoek, T. A., Vale, R. D. & Tanenbaum, M. E. Dynamics of Translation of Single mRNA Molecules in Vivo. *Cell* **165,** 976–989 (2016).

270. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17,** 1–19 (2016).

271. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

272. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

273. Marco-Sola, S., Sammeth, M., Guig??, R. & Ribeca, P. The GEM mapper: Fast, accurate and versatile alignment by filtration. *Nat. Methods* **9,** 1185–1188 (2012).

274. Decoene, T., Peters, G., De Maeseneire, S. & De Mey, M. Toward predictable 5'UTRs in Saccharomyces cerevisiae: Development of a yUTR calculator. *ACS Synth. Biol.* (2018). doi:10.1021/acssynbio.7b00366

275. Douin, V. *et al.* Use and comparison of different internal ribosomal entry sites (IRES) in tricistronic retroviral vectors. *BMC Biotechnol.* **4,** 16 (2004).

276. Edwards, S. R. & Wandless, T. J. Dicistronic regulation of fluorescent proteins in the budding yeast Saccharomyces cerevisiae Sarah. *Yeast* **27,** 229–236 (2009).

277. De Felipe, P. *et al.* E unum pluribus: Multiple proteins from a self-processing polyprotein. *Trends Biotechnol.* **24,** 68–75 (2006).

278. De Felipe, P., Hughes, L. E., Ryan, M. D. & Brown, J. D. Co-translational, intraribosomal cleavage of polypeptides by the foot-and-mouth disease virus 2A peptide. *J. Biol. Chem.* **278,** 11441–11448 (2003).

279. Szymczak, A. L. *et al.* Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide-based retroviral vector. *Nat. Biotechnol.* **22,** 589–594 (2004).

280. Kim, J. H. *et al.* High cleavage efficiency of a 2A peptide derived from porcine teschovirus-1 in human cell lines, zebrafish and mice. *PLoS One* **6,** 1–8 (2011).

281. Szymczak-Workman, A. L., Vignali, K. M. & Vignali, D. A. A. Design and construction of 2A peptide-linked multicistronic vectors. *Cold Spring Harb. Protoc.* **7,** 199–204 (2012).

282. Minskaia, E. & Ryan, M. D. Protein coexpression using FMDV 2A: Effect of 'linker' residues. *Biomed Res. Int.* **2013,** 12 pages (2013).

283. Gao, Z. liang, Zhou, J. hua, Zhang, J., Ding, Y. zhong & Liu, Y. sheng. The silent point mutations at the cleavage site of 2A/2B have no effect on the self-cleavage activity of 2A of foot-and-mouth disease virus. *Infect. Genet. Evol.* **28,** 101–106 (2014).

284. Unkles, S. E., Valiante, V., Mattern, D. J. & Brakhage, A. A. Synthetic biology tools for bioprospecting of natural products in eukaryotes. *Chem. Biol.* **21,** 502–508 (2014).

285. Wang, Y., Wang, F., Wang, R., Zhao, P. & Xia, Q. 2A self-cleaving peptide-based multi-gene expression system in the silkworm Bombyx mori. *Sci. Rep.* **5,** 16273 (2015).

286. Geier, M., Fauland, P., Vogl, T. & Glieder, A. Compact multi-enzyme pathways in P. pastoris. *Chem. Commun. (Camb).* **62,** 1643–1646 (2014).

287. Wang, S., Yao, Q., Tao, J., Qiao, Y. & Zhang, Z. Co-ordinate expression of glycine betaine synthesis genes linked by the FMDV 2A region in a single open reading frame in Pichia pastoris. *Appl. Microbiol. Biotechnol.* **77,** 891–899 (2007).

288. Beekwilder, J. *et al.* Polycistronic expression of a beta-carotene biosynthetic pathway in Saccharomyces cerevisiae coupled to beta-ionone production. *J. Biotechnol.* **192,** 383–392 (2014).

289. Park, M. *et al.* Expression of serotonin derivative synthetic genes on a single self-processing polypeptide and the production of serotonin derivatives in microbes. *Appl. Microbiol. Biotechnol.* **81,** 43–49 (2008).

290. Kuijpers, N. G. a *et al.* A versatile, efficient strategy for assembly of multi-fragment expression vectors in Saccharomyces cerevisiae using 60 bp synthetic recombination sequences. *Microb. Cell Fact.* **12,** 47 (2013).

291. Sharma, P. *et al.* 2A peptides provide distinct solutions to driving stop-carry on translational recoding. *Nucleic Acids Res.* **40,** 3143–3151 (2012).

292. Kuzmich, A. I., Vvedenskiĭ, A. V, Kopantsev, E. P. & Vinogradova, T. V. Quantitative comparison of expression for genes linked in bicistronic vectors via ires or 2A-peptide of porcine teschovirus-1 sequence. *Bioorg. Khim.* **39,** 454–65 (2013).

293. Berg JM, Stryer L. & Tymoczko, J. in *Biochemistry. 5th edition. New York: W H Freeman* (2002).

294. Liu, Z. *et al.* Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector. *Sci. Rep.* **7,** 2193 (2017).

295. Lee, S., Lim, W. a. & Thorn, K. S. Improved Blue, Green, and Red Fluorescent Protein Tagging Vectors for S. cerevisiae. *PLoS One* **8,** e67902 (2013).

296. Luke, G. A., Escuin, H., Felipe, P. De & Ryan, M. D. 2A to the Fore – Research , Technology and Applications. *Biotechnol. Genet. Eng. Rev.* **26,** 223–260 (2009).

297. Jakociunas, T. *et al.* CasEMBLR: Cas9-facilitated multi-loci genomic integration of in vivo assembled DNA parts in Saccharomyces cerevisiae. *ACS Synth. Biol.* **4,** 1226–34 (2015).

298. Horwitz, A. A. *et al.* Efficient Multiplexed Integration of Synergistic Alleles and Metabolic Pathways in Yeasts via CRISPR-Cas. *Cell Syst.* **1,** 88–96 (2015).

299. Schuetze, T. & Meyer, V. Polycistronic gene expression in Aspergillus niger. *Microb. Cell Fact.* **16,** 162 (2017).

300. Fernandes, L. D., de Moura, A. & Ciandrini, L. Gene length as a regulator for ribosome recruitment and protein synthesis: theoretical insights. *BioRxiv* 1–12 (2017). doi:10.1101/105296

301. Arava, Y. *et al.* Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 3889–3894 (2003).

302. Neymotin, B., Ettore, V. & Gresham, D. Multiple Transcript Properties Related to Translation Affect mRNA Degradation Rates in Saccharomyces cerevisiae. *Genes|Genomes|Genetics* **6,** 3475–3483 (2016).

303. Geisberg, J. V., Moqtaderi, Z., Fan, X., Ozsolak, F. & Struhl, K. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* **156,** 812–824 (2014).

304. Bonnin, P., Kern, N., Young, N. T., Stansfield, I. & Romano, M. C. Novel mRNA-specific effects of ribosome drop-off on translation rate and polysome profile. *PLoS Comput. Biol.* **13,** e1005555 (2017).

305. Hou, X. & Cheng, W. Detection of single fluorescent proteins inside eukaryotic cells using two-photon fluorescence. *Biomed. Opt. Express* **3,** 340 (2012).

306. Shashkova, S. & Leake, M. C. Single-molecule fluorescence microscopy review: shedding new light on old problems. *Biosci. Rep.* **0,** BSR20170031 (2017).

307. Partow, S., Siewers, V., Bjørn, S., Nielsen, J. & Maury, J. Characterization of different promoters for designing a new expression vector in Saccharomyces cerevisiae. *Yeast* **27,** 955–964 (2010).

308. Yuan, J. & Ching, C. B. Combinatorial Assembly of Large Biochemical Pathways into Yeast Chromosomes for Improved Production of Value-added Compounds. *ACS Synth. Biol.* **4,** 23–31 (2015).

309. Boerjan, W., Ralph, J. & Baucher, M. Lignin Biosynthesis. *Annu. Rev. Plant Biol.* **54,** 519–546 (2003).

310. Grotewold, E. *The Science of Flavonoids*. *The Science of Flavonoids* (2006). doi:10.1007/978-0-387-28822-2

311. Fraser, C. M. & Chapple, C. The phenylpropanoid pathway in Arabidopsis. *Arab. B.* **9,** e0152 (2011).

312. Braus, G. H. Aromatic amino acid biosynthesis in the yeast Saccharomyces cerevisiae: a model system for the regulation of a eukaryotic biosynthetic pathway. *Microbiol. Rev.* **55,** 349–70 (1991).

313. Prado, R., Angelica, E., Nielsen, J. & Borodina, I. Development of a yeast cell factory for production of aromatic secondary metabolites (PhD thesis). (2016).

314. Schnappauf, G., Krappmann, S. & Braus, G. H. Tyrosine and tryptophan act through the same binding site at the dimer interface of yeast chorismate mutase. *J. Biol. Chem.* **273,** 17012–17017 (1998).

315. Shirra, M. K. *et al.* Inhibition of Acetyl Coenzyme A Carboxylase Activity Restores Expression of the INO1 Gene in a snf1 Mutant Strain of Saccharomyces cerevisiae Inhibition of Acetyl Coenzyme A Carboxylase Activity Restores Expression of the INO1 Gene in a snf1 Mutant Strain. *Society* **21,** 5710–5722 (2001).

316. Zhang, M., Galdieri, L. & Vancura, A. The Yeast AMPK Homolog SNF1 Regulates Acetyl Coenzyme A Homeostasis and Histone Acetylation. *Mol. Cell. Biol.* **33,** 4701–4717 (2013).

317. Krivoruchko, A., Zhang, Y., Siewers, V., Chen, Y. & Nielsen, J. Microbial acetyl-CoA metabolism and metabolic engineering. *Metab. Eng.* **28,** 28–42 (2015).

318. Gaspar, P., Oliveira, J. L., Frommlet, J., Santos, M. A. S. & Moura, G. EuGene: Maximizing synthetic gene design for heterologous expression. *Bioinformatics* **28,** 2683–2684 (2012).

319. Gueldener, U., Heinisch, J., Koehler, G. J., Voss, D. & Hegemann, J. H. A second set of loxP marker cassettes for Cre-mediated multiple gene knockouts in budding yeast. *Nucleic Acids Res.* **30,** e23 (2002).

320. Agmon, N. *et al.* Yeast Golden Gate (yGG) for efficient assembly of S. cerevisiae transcription units. *ACS Synth. Biol.* **4,** 853–859 (2015).

321. Hartzog, P. E., Nicholson, B. P. & McCusker, J. H. Cytosine deaminase MX cassettes as positive/negative selectable markers in Saccharomyces cerevisiae. *Yeast* **22,** 789–798 (2005).

322. Güldener, U., Heck, S., Fiedler, T., Beinhauer, J. & Hegemann, J. H. A new efficient gene disruption cassette for repeated use in budding yeast. *Nucleic Acids Res.* **24,** 2519–2524 (1996).

323. Ronda, C. *et al.* Accelerating genome editing in CHO cells using CRISPR Cas9 and CRISPy, a web-based target finding tool. *Biotechnol. Bioeng.* **111,** 1604–1616 (2014).

324. Ralston, L., Subramanian, S., Matsuno, M. & Yu, O. Partial Reconstruction of Flavonoid and Isoflavonoid Biosynthesis in Yeast Using Soybean Type I and Type II Chalcone Isomerases. *Plant Physiol.* **137,** 1375–1388 (2005).

325. VanderMolen, K. M., Raja, H. A., El-Elimat, T. & Oberlies, N. H. Evaluation of culture media for the production of secondary metabolites in a natural products screening program. *AMB Express* **3,** 71 (2013).

326. Yan, Y., Kohli, A. & Koffas, M. A. G. Biosynthesis of natural flavanones in Saccharomyces cerevisiae. *Appl. Environ. Microbiol.* **71,** 5610–5613 (2005).

327. Leonard, E. & Koffas, M. a G. Engineering of artificial plant cytochrome P450 enzymes for synthesis of isoflavones by Escherichia coli. *Appl. Environ. Microbiol.* **73,** 7246–7251 (2007).

328. Stahlhut, S. G. *et al.* Assembly of a novel biosynthetic pathway for production of the plant flavonoid fisetin in Escherichia coli. *Metab. Eng.* **31,** 84–93 (2015).

329. Helmstaedt, K., Strittmatter, A., Lipscomb, W. N. & Braus, G. H. Evolution of 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase-encoding genes in the yeast Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci.* **102,** 9784–9789 (2005).

330. Natarajan, K. *et al.* Transcriptional Profiling Shows that Gcn4p Is a Master Regulator of Gene Expression during Amino Acid Starvation in Yeast Transcriptional Profiling Shows that Gcn4p Is a Master Regulator of Gene Expression during Amino Acid Starvation in Yeast. *Mol. Cell. Biol.* **21,** 4347–4368 (2001).

331. Strucko, T., Magdenoska, O. & Mortensen, U. H. Benchmarking two commonly used Saccharomyces cerevisiae strains for heterologous vanillin-beta-glucoside production. *Metab. Eng. Commun.* **2,** 99–108 (2015).

332. Alam, M. T. *et al.* The metabolic background is a global player in Saccharomyces gene expression epistasis. *Nat. Microbiol.* **1,** 15030 (2016).

333. Paciello, L., Zueco, J. & Landi, C. On the fermentative behavior of auxotrophic strains of Saccharomyces cerevisiae. *Electron. J. Biotechnol.* **17,** 246–249 (2014).

334. Rodriguez, A. *et al.* Comparison of the metabolic response to over-production of p-coumaric acid in two yeast strains. *Metab. Eng.* **44,** 265–272 (2017).

335. Jez, J. M., Ferrer, J. L., Bowman, M. E., Dixon, R. A. & Noel, J. P. Dissection of malonyl-coenzyme A decarboxylation from polyketide formation in the reaction mechanism of a plant polyketide synthase. *Biochemistry* **39,** 890–902 (2000).

336. Costa, M. A. *et al.* Characterization in vitro and in vivo of the putative multigene 4-coumarate:CoA ligase network in Arabidopsis: Syringyl lignin and sinapate/sinapyl alcohol derivative formation. *Phytochemistry* **66,** 2072–2091 (2005).

337. Trantas, E., Panopoulos, N. & Ververidis, F. Metabolic engineering of the complete pathway leading to heterologous biosynthesis of various flavonoids and stilbenoids in Saccharomyces cerevisiae. *Metab. Eng.* **11,** 355–366 (2009).

338. Liu, W., Zhang, B. & Jiang, R. Improving acetyl-CoA biosynthesis in Saccharomyces cerevisiae via the overexpression of pantothenate kinase and PDH bypass. *Biotechnol. Biofuels* **10,** 41 (2017).

339. Vos, T., de la Torre Cortés, P., van Gulik, W. M., Pronk, J. T. & Daran-Lapujade, P. Growth-rate dependency of de novo resveratrol production in chemostat cultures of an engineered Saccharomyces cerevisiae strain. *Microb. Cell Fact.* **14,** 133 (2015).

340. Lehka, B. J. *et al.* Improving heterologous production of phenylpropanoids in Saccharomyces cerevisiae by tackling an unwanted side reaction of Tsc13, an endogenous double-bond reductase. *FEMS Yeast Res.* **17,** 1–12 (2017).

341. Kohlwein, S. D. *et al.* Tsc13p is required for fatty acid elongation and localizes to a novel structure at the nuclear-vacuolar interface in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **21,** 109–25 (2001).

342. Venayak, N., Anesiadis, N., Cluett, W. R. & Mahadevan, R. Engineering metabolism through dynamic control. *Curr. Opin. Biotechnol.* **34,** 142–152 (2015).

343. Weinhandl, K., Winkler, M., Glieder, A. & Camattari, A. Carbon source dependent promoters in yeasts. *Microb. Cell Fact.* **13,** 5 (2014).

344. Hughes, R. A. & Ellington, A. D. Synthetic DNA Synthesis and Assembly : Putting the Synthetic in Synthetic Biology. *Cold Spring Harb Perspect Biol* **9,** 1–17 (2017).

345. Annaluru, N. *et al.* Total Synthesis of a Functional Designer Eukaryotic Chromosome. *Science (80-.).* **344,** 55–58 (2014).

346. Richardson, S. M. *et al.* Design of a synthetic yeast genome. *Science (80-.).* **355,** 1040–1044 (2017).

347. Parekh, S., Vinci, V. A. & Strobel, R. J. Improvement of microbial strains and fermentation processes. *Appl. Microbiol. Biotechnol.* **54,** 287–301 (2000).

348. Rugbjerg, P., Myling-Petersen, N., Porse, A., Sarup-Lytzen, K. & Sommer, M. O. A. Diverse genetic error modes constrain large-scale bio-based production. *Nat. Commun.* **9,** 787 (2018).

SUMMARY

In the last decades, the industrial or white biotechnology which uses micro-organisms and enzymes for the sustainable production of industrial relevant compounds is on the rise. One group of such interesting compounds are flavonoids, plant secondary metabolites with promising bioactive properties for treatments against viral and bacterial infections, cancer and inflammation. As such, these molecules attain huge attention for usage in the human health sector making their secured and defined supply essential. Currently, the industrial production of these molecules has some inherent drawbacks like low yields using plant extraction and multiple reaction steps, harsh reaction conditions and the difficulty of chiral centers in chemical synthesis. As such, production of these specialized plant metabolites in microbial cell factories can be a valuable alternative. However, developing suitable microbial strains with profitable product titers for an industrial environment is challenging, especially due to the difficulty of tuning all steps in a (heterologous) production pathway and the native metabolism. In this respect, tremendous efforts to enhance the strain development process in the field of metabolic engineering and synthetic biology have been made. Nevertheless, this still remains a cost – and labor-intensive undertaking, not the least in the attractive eukaryotic host *Saccharomyces cerevisiae*. To this end, this doctoral research aimed to develop novel tools to facilitate the alteration of gene expression at the transcriptional and translational level, as such speeding up the construction of yeast cell factories which was applied here on a naringenin production strain as proof of concept.

The use of characterized, modular regulatory parts and the standardized sharing of biological data plays an indispensable role to transform the synthetic biology field to a mature engineering field. In this respect, the obscure demarcation of yeast's transcriptional and translational control elements slows down this transformation process in eukaryotes. As such, novel biological parts on the one hand influencing transcription, *i.e.* semi-synthetic core promoters, and on the other hand affecting translation, *i.e.* 5'UTRs with a predictive outcome on gene expression, were developed. The yeast core promoter is known to be the main determinant of transcription levels, making it an interesting target for modifying biosynthetic pathways. Additionally, minimal core promoters with equal or better activities as the cumbersome native yeast promoters could immensely facilitate the assembly of transcription units. Therefore, the well-characterized *TEF1* promoter was truncated to elucidate the minimal length needed for functional gene expression. This minimal sequence served as template for the creation of a core promoter library leading to short, functional semi-synthetic core promoters which were equally or twice as strong as commonly long

yeast promoters. Besides modulating transcription, altering a gene's translation initiation rate has proven to work well as a tool to predictably modify gene expression, especially in prokaryotes. Therefore, a similar forward engineering approach was set up in *S. cerevisiae* by developing a partial least square (PLS) regression model linking 13 5'UTR features with protein levels. This model was used for the *de novo* design of 5'UTRs with a predictive outcome on gene expression in different genetic contexts. *In vivo* testing of these 5'UTR sequences showed a good general applicability of the model since adequate coefficients of determination (R^2) were obtained in all experiments. As such, this data-driven algorithm expands the small toolbox of existing methods for the novel design of biological parts in yeast.

Besides monocistronic regulation, previous studies have shown the possibility of eukaryotic pathway balancing through multicistronic expression. However, this is still a mainly unexplored tool in *S. cerevisiae*. To this end, a thorough evaluation of this technique was performed by the usage of T2A peptides enabling ribosome skipping at the end of a coding sequence and proven to be efficient in different yeast species. Typically, their multiple use in long pathways is hindered because of the risk of unwanted homologous recombination. To allow this, five T2A sequences were developed differing as much as possible in their nucleotide sequence and evaluated for their effectiveness as a tool for pathway optimization. The T2A peptides, with the exception of one T2A having some lower reliability, effectively led to spliced proteins. Finally, their performance as real regulatory elements in a polycistronic pathway was tested in the genome for bi-, tri-, and quadcistronic constructs. While all constructs were stably integrated in the genome and for bi and tricistronic expression acceptable protein levels was observed, a complete lack of expression was noticed for the last positioned protein in the quadcistronic transcription unit. To this end, the usage of multicistronic pathways in baker's yeast is preferably limited to bi- and tricistronic expression units.

To show the ability of *S. cerevisiae* as an industrial host for the biosynthesis of specialty metabolites, the S288c wild-type yeast was transformed into a cell factory for naringenin production. To do so, cutting-edge synthetic biology tools such as CRISPR/Cas9 and the versatile genetic assembly system (VEGAS) were used. Yeast's native metabolism was rewired to enhance the supply of flavonoid precursors phenylalanine, tyrosine and malonyl-CoA. First, the improvement of the phenylalanine and tyrosine pool was assessed indirectly

by measuring p-coumaric acid after introduction of its pathway. Next, the augmented cytosolic malonyl-CoA pool was evaluated by analyzing naringenin titers by introducing the last three genes of the pathway and feeding the strains with p-coumaric acid. Finally, both approaches were combined in the strains with the most promising metabolic backgrounds to produce naringenin *de novo* from glucose with maximal productivity. Acceptable titers up to 4.0 mg/l were obtained. However, to reach a full profitable production strain further optimization will be needed. As only the native precursor pools were modified and no fine-tuning of the naringenin pathway itself was performed yet, the latter looks the most obvious way to be working on as a future perspective to increase final product titers. To this end, our developed design tool for 5'UTRs was tested for the predictive expression of the *Rhodobacter capsulatus tal1* gene, converting tyrosine to p-coumaric acid. The initial results were promising for further pathway optimization in that way that p-coumaric acid titers were proportional with the predicted protein abundance.

In general, several tools were developed and evaluated during this Ph.D. research which could facilitate future development and optimization of *S. cerevisiae* cell factories. More specifically, short semi-synthetic core promoters and a forward engineering approach to alter a gene's translation were constructed. Also the capacity of 2A peptides as a tool for multicistronic expression in yeast was investigated. Additionally, since the main goal of industrial biotechnology is to set up green production processes for economically relevant compounds, a naringenin producing yeast strain was created. Overall, this Ph.D. dissertation showed the potential of *S. cerevisiae* as an interesting host for secondary metabolite production and contributed to the expansion of the yeast synthetic biology toolbox enabling the reduction of strain development times for future bioprocesses.

SAMENVATTING

De industriële of witte biotechnologie die gebruik maakt van micro-organismen en enzymen voor de productie van industrieel relevante verbindingen is de laatste jaren aan een opmars bezig. Een interessante groep van verbindingen hiertoe zijn flavonoïden, secundaire plantmetabolieten met veelbelovende bioactieve eigenschappen voor behandelingen tegen virale en bacteriële infecties, kanker en ontstekingen. Bijgevolg krijgen deze moleculen bijzondere aandacht voor hun gebruik in de gezondheidssector. Het is dus essentieel om deze componenten op een duurzame manier en in voldoende hoeveelheden te voorzien. De huidige productie van deze moleculen heeft enkele inherente nadelen zoals de lage opbrengst na extractie uit planten en de meerdere reactiestappen, brute reactiecondities en de bijkomende moeilijkheid van chirale centra bij chemische synthese. De productie van deze hoogwaardige secundaire metabolieten met behulp van micro-organismen is dan ook een valabel alternatief. Echter, de ontwikkeling van geschikte microbiële stammen met een rendabele productietiter is een hele uitdaging, zeker door de moeilijkheid van het afstemmen van alle stappen in een (heterologe) *pathway* en het natieve metabolisme. In dit opzicht hebben recente technieken uit het veld van de synthetische biologie en *metabolic engineering* er wel voor gezorgd dat dit proces deels vereenvoudigd werd. Desondanks blijft het nog altijd een kost – en arbeidsintensieve onderneming, en niet in het minst in het eukaryoot gastheerorganisme *Saccharomyces cerevisiae*. Daarom is het doel van dit doctoraatsonderzoek om nieuwe technologieën te ontwikkelen die het wijzigen van genexpressie op transcriptie – en translatieniveau vergemakkelijkt en als dusdanig de constructie van gist productiestammen versnelt. Met het belang van flavonoïden voor de gezondheidssector werd dit toegepast voor de productie van naringenine als *proof of concept*.

Het gebruik van gekarakteriseerde, modulaire regulerende DNA sequenties en het delen van biologische data op een gestandaardiseerde manier zijn essentieel voor de transformatie van het synthetische biologie veld naar een volwaardige *engineering* discipline. In dat opzicht vormt de vage afbakening van transcriptionele en translationele controle elementen in gist een hinderpaal om dit transformatieproces in eukaryoten te versnellen. Om dit aan te pakken werden nieuwe biologische regulatoren ontwikkeld die enerzijds een effect hadden op transcriptie, *i.e.* semisynthetische *core* promotoren, en anderzijds een effect hadden op translatie, *i.e.* 5'UTRs met een voorspelbare invloed op genexpressie. De *core* promotor in gist heeft de grootste invloed op de transcriptionele modulatie van een gen waardoor het een interessante target is om bio-synthetische

pathways te reguleren. Daarnaast kunnen minimale *core* promotoren voor gist met gelijke of betere activiteiten dan de bestaande logge natieve promotoren, het assembleren van transcriptie eenheden sterk vergemakkelijken. Bijgevolg werd de goed gekende *TEF1* promotor ingekort om de minimale lengte te achterhalen die aanleiding gaf tot voldoende genexpressie. Deze minimale sequentie deed verder dienst als template voor het creëren van een *core* promotor bank die verder resulteerde in korte, semisynthetische promotoren die even of zelfs dubbel zo sterk waren als de veelgebruikte lange gist promotoren. Naast het variëren van transcriptie heeft het aanpassen van de translatie-initiatie, voornamelijk in prokaryoten, reeds zijn nut bewezen als technologie om op een voorspelbare manier genexpressie te modificeren. Een gelijkaardige methode werd daarom opgezet in *S. cerevisiae* door de ontwikkeling van een *partial least square* (PLS) regressiemodel dat 13 5'UTR eigenschappen linkt met eiwitniveaus. Dit model werd verder gebruikt voor *de novo* design van 5'UTRs met een voorspelbaar effect op genexpressie in verschillende genetische contexten. *In vivo* evaluatie van deze 5'UTR sequenties toonde de goeie algemene toepasbaarheid van het model aan, aangezien adequate determinatie coëfficiënten (R^2) bekomen werden in alle experimenten. Dit data-gedreven algoritme draagt zo bij tot de uitbreiding van de eerder beperkte set van bestaande methoden voor *de novo* design van biologische regulatoren in gist.

Naast monocistronische regulatie toonden eerdere studies reeds de mogelijkheid aan van multicistronische expressie voor het balanceren van eukaryotische *pathways*, echter, deze techniek wordt grotendeels niet gebruikt in *S. cerevisiae*. Daarom werd deze technologie, gebruik makende van T2A peptide sequenties die zorgen voor ribosoom *skipping* op het einde van een coderende sequentie en waarvan aangetoond werd dat ze efficiënt werken in verschillende gist species, grondig geëvalueerd. Het herhaaldelijk gebruik in lange *pathways* wordt echter verhinderd in bakkersgist door de hoge kans op ongewenste homologe recombinatie. Om dit te vermijden werden vijf T2A sequenties gemaakt die zoveel mogelijk verschilden in hun onderlinge nucleotide sequentie en verder werden deze geëvalueerd op hun effectiviteit als een regulerend element voor *pathway* optimalisatie. De T2A peptiden, met uitzondering van één T2A peptide met een wat lagere betrouwbaarheid, leidden effectief tot gesplitste eiwitten. Finaal werd voor bi-, tri- en quadcistronische constructen in het genoom getest of 2A peptiden dienst kunnen doen als een echte regulator van genexpressie in een polycistronische *pathway*. Terwijl alle constructen stabiel geïntegreerd werden in het genoom en voor bi – en tricistronische expressie acceptabele eiwitniveaus

vastgesteld werden, werd er geen expressie waargenomen voor het laatst gepositioneerde eiwit in de quadcistronische transcriptie-eenheid. Bijgevolg kan geconcludeerd worden dat het gebruik van multicistronische expressie in bakkersgist best beperkt blijft tot bi- en tricistronische expressie.

Om het vermogen aan te tonen van *S. cerevisiae* als een industriële gastheer voor de biosynthese van hoogwaardige metabolieten werd het S288c wild-type getransformeerd in een productiestam voor naringenine. Om dit te verwezenlijken werd gebruik gemaakt van baanbrekende technieken uit de synthetische biologie zoals CRISPR/Cas9 en het veelzijdig genetische assemblage systeem VEGAS. Het metabolisme van gist werd omgebouwd om de aanvoer van de flavonoïde precursoren fenylalanine, tyrosine en malonyl-CoA te verhogen. Als eerste werd de verhoogde *pool* aan fenylalanine en tyrosine indirect geanalyseerd door coumarinezuur te meten na introductie van de coumarinezuur *pathway*. Daarna werd de verbeterde *pool* aan malonyl-CoA in het cytosol geëvalueerd door het analyseren van naringenine titers na introductie van de laatste drie naringenine *pathway* genen en de stammen te voeden met coumarinezuur. Finaal werden beide optimalisatie methoden gecombineerd in de stammen met de meest belovende metabolische achtergrond om naringenine te produceren vanuit glucose met maximale productiviteit. Aanvaardbare titers tot 4.0 mg/l werden bekomen, echter, om een volledig rendabele productiestam te bekomen zal verdere stamoptimalisatie nodig zijn. Aangezien enkel de natieve precursor *pools* gemodificeerd werden en nog geen afstelling van de naringenine *pathway* zelf uitgevoerd werd, lijkt dit laatste de meest aangewezen weg om in de toekomst de finale productietiter te verhogen. Daartoe werd de ontwikkelde design-methode voor 5'UTRs uitgetest voor de voorspelbare expressie van het *Rhodobacter capsulatus tal1* gen, dat instaat voor de conversie van tyrosine naar coumarinezuur. De eerste resultaten waren veelbelovend aangezien de coumarinezuur-titers proportioneel waren met de voorspelde enzymniveaus, wat nogmaals het potentieel aantoont van deze technologie voor verdere *pathway* optimalisatie.

Algemeen werden tijdens dit doctoraatsonderzoek verscheidene technologieën ontwikkeld en geëvalueerd om de toekomstige ontwikkeling en optimalisatie van *S. cerevisiae* productiestammen te vergemakkelijken. Specifiek werden korte semisynthetische *core* promotoren ontwikkeld samen met een methode om betrouwbaar de translatie van een gen te wijzigen. Daarnaast werd ook de capaciteit nagegaan van 2A peptiden als een regulator

voor multicistronische expressie in gist. Aangezien het ontwikkelen van duurzame productieprocessen voor economisch relevante verbindingen een belangrijke missie is van de industriële biotechnologie, werd ook een naringenine productiestam gecreëerd. In zijn geheel heeft dit Ph.D. onderzoek aangetoond dat *S. cerevisiae* een interessante gastheer is voor de productie van secundaire metabolieten en heeft het bijgedragen tot de uitbreiding van synthetische biologie technologieën voor gist. Dit zal in de toekomst bijdragen bij het verder reduceren van ontwikkelingstijden van nieuwe bioprocessen.

CURRICULUM VITAE

Thomas Decoene (°June 15, 1990)

Education

2014 – 2017	**Doctor of Applied Biological Sciences**, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium Ph.D. thesis: *Expanding the portfolio of synthetic biology tools in Saccharomyces cerevisiae for the optimization of heterologous production pathways at the transcriptional and translational level* Promoters: Prof. Marjan De Mey (Ghent University) and dr. Sofie De Maeseneire (Ghent University)
2011 – 2013	**Master of Science in Bioscience Engineering: Chemistry and Bioprocess Technology**, Ghent University, Ghent, Belgium Master thesis: *Saccharomyces cerevisiae as biotechnological production platform organism: a myo-inositol case study* Promoters: Prof. Marjan De Mey (Ghent University) and dr. Sofie De Maeseneire (Ghent University)
2008 – 2011	**Bachelor of Science in Bioscience Engineering: Chemistry and Food Technology**, Ghent University, Ghent, Belgium Bachelor thesis: *A critical evaluation of dust in harbor industry* Promoter: Prof. Paul Van der Meeren
2002 – 2008	Science and mathematics, Broederschool, Roeselare, Belgium

Specialized courses

2016	**Data manipulation, analysis and visualization in Python** Ghent University, Ghent, Belgium
2016	**qPCR-course – Biogazelle** Ghent University, Ghent, Belgium
2014	**Nano3Bio workshop: Bioinformatics** Ghent University, Ghent, Belgium
2014	**Ghent Biobased Economy Summer School** Ghent University, Ghent, Belgium

Transferable skills seminars

2017	**Schrijven voor niet-vakgenoten en pers** Ghent University, Ghent, Belgium
2014	**7th From Ph.D. to Job Market** Arcelor Mittal, Ghent, Belgium
2014	**Introduction Day 2014 for new Ph.D's** Ghent University, Ghent, Belgium

Attended conferences

2017	**7th International meeting on Synthetic Biology (SB 7.0)**, June 13-16, Singapore, Singapore Poster presentation: Towards predictable 5'UTRs in *Saccharomyces cerevisiae*
2016	**Yeasterday 2016**, May 13, Leuven, Belgium Poster presentation: Protein expression modulation in *S. cerevisiae* through random assembly of an UAS library
2016	**2nd edition of Genome Engineering and Synthetic Biology: Tools and Technologies (GESB 2016)**, January 28-29, Ghent, Belgium Poster presentation: Engineering transcription in *S. cerevisiae* through random assembly of an UAS library
2015	**Enabling Technologies for Eukaryotic Synthetic Biology (EMBO-EMBL symposium)**, June 21-23, Heidelberg, Germany Poster presentation: Engineering eukaryotic transcription in yeast: towards the minimal core promoter
2014	**5th CINBIOS Forum for Industrial Biotechnology and the Biobased Economy**, November 07, Mechelen, Belgium Poster presentation: A balancing act: building the aurone pathway in *Saccharomyces cerevisiae*
2014	**Yeasterday 2014**, June 06, Utrecht, The Netherlands Poster presentation: Assembly and optimization of synthetic pathways in *Saccharomyces cerevisiae*: aurones as case study

Student guidance

Practical exercises General Microbiology (Bachelor of Science in Bioscience engineering, 2nd year):

- Academic year 2013-2014
- Academic year 2014-2015
- Academic year 2015-2016
- Academic year 2017-2018

Tutor in student theses

2015-2016	**Yatti De Nijs**, M.Sc.: Synthetische biologie *tools* voor de ontwikkeling van een flavonoïde precursorstam in *Saccharomyces cerevisiae*
2014-2015	**Nathalie Cuypers**, M.Sc.: Ontwikkeling van *engineering tools* en een flavonoïde precursorstam in *Saccharomyces cerevisiae*
2013-2014	**Jan De Kezel, Anne-Marie Jennen, Kevin Vandoorne & Bram Van Renterghem**, B.Sc.: Flavonoïden, een interessante target voor microbiële productie?

Accepted publications

Decoene, T., De Paepe, B., Maertens, J., Coussement, P., Peters, G., De Maeseneire, S. L., and De Mey, M. (2017) Standardization in synthetic biology: an engineering discipline coming of age. *Crit. Rev. Biotechnol. Sep 27*, 1–10. doi: 10.1080/07388551.2017.1380600

Decoene, T., Peters, G., De Maeseneire, S., & De Mey, M. (2018). Toward predictable 5'UTRs in *Saccharomyces cerevisiae*: Development of a yUTR calculator. *ACS Synthetic Biology*. https://doi.org/10.1021/acssynbio.7b00366

Submitted publications

Decoene, T., De Nijs Y., Delmulle, T., Cuypers, N., De Maeseneire, S. L., and De Mey, M. (2018). Engineering the native precursor pools of *Saccharomyces cerevisiae* for the sustainable production of flavonoids: a naringenin case study. *Biotechnology and Bioengineering*.